Vol. XXIV
No. 4

PSYCHOLOGICAL REVIEW PUBLICATIONS

Whole No. 105
1917

Psychological Monographs

STUDIES FROM THE PSYCHOLOGICAL LABORATORY OF THE UNIVERSITY OF CHICAGO

The Reliability of Mental Tests in the Division of an Academic Group

By

BEARDSLEY RUML, Ph.D.
Instructor in Psychology
Carnegie Institute of Technology

PSYCHOLOGICAL REVIEW COMPANY
PRINCETON, N. J.
AND LANCASTER, PA.

Agents: G. E. STECHERT & CO., London (2 Star Yard, Carey St., W. C.)
Leipzig (Koenigstr., 37); Paris (16 rue de Condé)

I

The object of this research is the determination of the value of mental tests in dividing large groups of students into smaller groups of relatively equal mental ability. Such a division of students may be desirable for two reasons; the groups may be merely indicated in order that administrative officers may direct and advise students with greater confidence, or the large mixed class may be actually broken into groups of greater intellectual homogeneity so that no class need contain individuals who differ greatly in intelligence. In institutions where several sections of the same course must be offered, the sections might be formed on the basis of the abilities of the students instead of by a division of the total group according to the first letter of the individuals' surnames.

To indicate more clearly the sort of heterogeneity in which we are interested, it may be desirable to mention several ways in which students may differ from one another. In the first place, individuals may differ in the amount and quality of their previous academic training. The school attempts to remove these differences by imposing entrance requirements, and by providing courses in many departments suited to various degrees of preparation. Secondly there are differences in interests. That most institutions do not consider these differences of great importance in the first year of college is seen in the fact that the work of the freshman year is largely prescribed. Allowance is made for differences in interests in a wealth of elective courses in the junior and senior years. Thirdly, individuals may vary in *persistence* independently of differences in interests. Persistence of motive, the common character factor for which Webb[26] argues, may be the basis for one type of individual variation. In the fourth place there may be differences in special abilities. It is not possible at present to estimate the importance of special ability *independent* of general ability and special interests; but the high average correlation found by Carey[4] between geography, science,

history and mathematics, +.74, indicates that special ability in any one of these subjects cannot be so important a factor as is popularly supposed. "The correlations between the various college subjects are all positive and argue against the commonly expressed belief in rather close specialization of abilities; the student who does well in one of these subjects tends to do well in all of them."[10] Occasional cases of rare special ability in particular college subjects will doubtless occur, but such cases are easily recognized.

Lastly, individuals may differ in general intelligence, or brightness. The remainder of this monograph will attempt to show that by the use of mental tests, classes of students relatively homogeneous in this respect may actually be selected from the mixed group. It will further discuss from the standpoint of accuracy the advantages of different divisions of the total group. Clearly the total group may be divided into any number of sub-groups, and these sub-groups may contain any percentage of the total group. The problem is to choose the most accurate method of division. And finally, some evidence of a novel sort will be presented showing the relation between performance in mental tests and ability.

II

I

The discussion of the mental tests to be used in the selection of homogeneous classes, and of the methods of treating the measurements must be postponed pending the selection of a criterion by which the reliability of such tests may be judged. There is no a priori reason for believing that a series of mental tests will give an accurate indication of the intelligence of the individual tested. It is not hard to understand why such a series of tests *might* give a good index of ability, but to assert that this is the fact would be sheer dogmatism. As a matter of fact, it is seriously questioned whether tests devised to measure certain particular functions, such as memory or attention, really give a fair representation of those activities, so that considerably more doubt may

be attached to the value of psychological tests as measures of the less well defined function, general intelligence. It seems imperative, therefore, to abandon once and for all, performance in all the tests combined as the basis for judging the accuracy of any one test, not because standing in all tests combined is an unreliable measure of intellectual ability, but because it is not known in advance to be actually reliable.

Since the mental tests are to be used in a practical situation, in the division of a mixed group into smaller groups of more homogeneous intellectual ability, it seems natural to choose as a criterion of their value the success with which they handle the concrete problem of making the separation. The evaluation of the tests then consists in giving them experimentally to several groups of freshmen, and in determining how accurately the group would have been divided had the tests actually been in use.

The problem of evaluating the tests with reference to this concrete situation is therefore seen to presuppose the solution of another problem, namely that of determining how the experimental groups which have been tested should have been divided. If there were sharp and discrete classes of intelligence, for example, *good, mediocre* and *poor,* in which individual abilities might be placed, there would be little difficulty in finding how the separation should have been made. But unfortunately, general intelligence does not lend itself to any such rigid classification. On the contrary it seems to proceed through a continuous range of variation from the very bright to the very dull. As a result, a relatively homogeneous class would consist of the best or worst ten, twenty or thirty percent of the entire group. Such considerations show that it is impossible to designate any *definite* percentage of individuals as bright or dull. And, therefore, before any division whatever can be made, the individuals in the experimental group must be arranged in the order of their abilities. In other words, since general intelligence does not naturally divide into clear-cut categories, it is necessary that a value of some sort be predicated of the ability of every individual.

The mental ability of the individuals in experimental groups may be approximated in several ways. We have attempted to

show that to use as a measure of this ability the performance in the combined tests is inconsistent when the tests themselves are the objects of investigation. There remains the possibility of using either the estimations of instructors or other indices of the actual achievements of the students, provided it can be shown that either of these measures will properly evaluate the abilities of the students.

Two limitations to the usefulness of the tests seem to follow from selecting either estimations or achievements as the criterion by which the reliability of the tests in our experimental groups is measured. Since the tests are evaluated high or low according to the resemblance between the orders of individual abilities as given by the tests and as given by the criterion, it seems logical to inquire at what point the need for tests arises; why not use the information given by whatever criterion is adopted, as the basis for securing homogeneity? It may be urged that if the tests are required only to duplicate the information given by some other method already available, the tests in reality add nothing and had best be neglected as an unnecessary encumbrance to an overfull administrative program.

There are two answers to this objection. If the resemblance between the order of individuals as given by tests and as given by the criterion were complete, then the possibility that the tests might give new information would be eliminated. But the following facts must be taken into consideration: no matter how carefully the criterion of ability is selected, the chances are that it will be slightly in error. Now if the tests give a fairly high approximation, but not a duplication, of the order given by the criterion, there is reason to suppose that in some of the cases where the criterion is wrong the tests may be right. This answer to the objection must be taken only as a suggestion, for it is based upon an appeal from the ultimate reliability of the criterion, and reliability to a high degree in the criterion is a necessary presupposition for its use. Secondly, the tests may be given to the individuals, and the homogeneous classes may be formed on the basis of test performance long before either of the other criteria is available. If marks obtained in college are used, surely the average of not less than a year would have to be taken; marks from preparatory schools cannot be used because they represent evaluations of achievement based upon different standards; marks on entrance examinations are subject to errors which will be discussed later. If estimations of ability are used as criteria, the judge should have the individual judged under observation for a period of from one term to a year. Many of the advantages of homogeneous groups come in the earliest portions of college work, in exactly the periods where neither judgments nor grades can be obtained. Thus the value of mental tests will consist not so much in giving information that could not be obtained otherwise, but in giving the information *immediately,* so that the selection of groups may be made when selection is of the greatest value.

A second objection to the use of either instructors' judgments of ability or students' achievements is that these measures give no indication of the

individual's *general ability*. A business man might judge an individual in a very different manner from an instructor, and after all, which is the better able to define general intelligence? The student who secures the highest grades in classroom work may have very little ability in handling the affairs of practical everyday life. Is it possible to say that tests which correlate with criteria of the academic kind are for that reason measures of ability?

The person who makes an objection of this kind will not be content with the statement that students who secure good grades may not be interested in practical affairs; nor will the observation that the teacher's business is pretty largely concerned with the reactions of the mind, convince him that the teacher is a good judge of intelligence. And so the objection will not be met; but we shall so specify the ability under discussion that this critic may not take offense. If our tests agree with the academic criteria of ability, we shall claim only that tests may be used to select groups of homogeneous *academic* ability. This is sufficient for our purpose, since the use of mental tests in the academic situation is the object of our study. Now although instructors may not be capable of judging individuals in a way that would satisfy all standards of intelligence, they are certainly the most acceptable source of information as to the abilities of the students in academic work. If the reader is inclined to accept the academic criterion as the basis of approximating intelligence in general, so much the better; if mental tests are found to be accurate here, their applicability will be just that much wider.

One, or perhaps both, of the academic criteria may then be used to tell how the experimental groups should have been divided. There remains the examination of the relative advantages of grades and of estimations.

2

Grades, or credits, have had a wide use as measures of the ability of students. Wissler,[27] in early work at Columbia, selected grades as his criterion; and in the most recent reports of test work on college students at hand, that of Rowland and Lowden[17] at Reed College, and of Bell[1] at the University of Texas grades are still favored. Grades, aside from being records of the actual academic *achievement* of students, have several important advantages which probably account for their popularity as an index of ability. In the first place, they are easily obtained. It is necessary that the instructor turn into the administrative office a mark showing the performance of the student in his course, and this mark once turned in is ever afterward available for the purposes of research. This availability of grades is a very important virtue, and must be seriously con-

sidered. In the second place, grades lend themselves easily to averaging. If the grade is expressed numerically the average of the student's abilities in any number of courses may be expressed directly; if the grade is expressed by letters, the institution may have some credit scheme whereby a numerical value may be attached to each letter. Averaging makes possible a partial neutralization of accidental errors, and thus greatly increases the reliability of the measure. In the third place, grades usually indicate that there are a few individuals who are very good, and a few others who are very poor, with the majority clustered about some mid-value. This is just the kind of distribution that would be expected in the case of abilities, and it takes into account the fact that the difference between two adjacent individuals at either of the extremes of ability is greater than the difference between two individuals near the mean ability. Finally, assuming that the standard of grading does not change greatly from year to year, we may combine into a single group individuals who were in reality members of successive classes. This advantage is important since it makes possible the formation of a single group of any desired size.

In spite of these many advantages, the use of grades as a criterion of academic ability is open to serious objections. Too often instructors make of the grade an administrative device, inciting certain students to greater efforts. To bright students a lower mark may be given than is actually deserved; to poor students a higher mark may be given, just as an encouragement. In these cases the grade has a complex meaning; it is no longer a simple measure of academic achievement.

Then too, the precise ability represented by any mark is not defined, so that instructors grade by very different standards. As a result, all students must be graded by exactly the same instructors, or else the grades given by all instructors must be stated in terms of the same average and same dispersion. Otherwise, the taking of an average is not allowable, and the grade criterion instantly loses one of its greatest advantages.

Finally, grades as measures of ability are subject to both accidental and constant errors. The difference between constant er-

rors and accidental errors must be carefully noted, for it will play a very important part in discussions to follow.[8] An accidental error is an error that is produced by the interplay of many irregular and unrelated influences, and as a result the empirical measurement may be either greater or less than the true value of the measured object. As long as errors are accidental, it is possible to obtain a close approximation to the true value by determining the point where the mean square deviation of the empirical measurements is a minimum (the arithmetic mean of the measurements) and by using this point as the true value.[14] The several grades that are obtained by a student may be considered empirical measurements of academic achievement, and as such they are subject to many accidental errors. These are due to such causes as variable personal reactions of instructors, and variable interests of the student.

But grades are also affected by many *constant errors.* These are errors which tend to displace an individual's total standing, since they tend to displace each of the student's grades in the same direction. Severe economic pressure, social interests, general ill health or temporary absences, tendency to nervous instability at times of recitation or examination,—these factors operate on all grades in precisely the same way. For example, if a student gives an undue amount of time to social activities, *all* of his grades will be lowered, *and no amount of averaging will even approximate a true index of his ability.* In the physical sciences, the standard method of eliminating the effect of constant errors is to determine their magnitude, and to make the necessary corrections in the measurements. Corrections for the effect of temperature and barometric pressure are readily made, but the problem of estimating how much lower a student's grades are because of the fact that he traveled for a week with the dramatic club is insoluble.

There is considerable variability in the amount of constant error in different institutions, depending upon the attitude of the student body toward the work of the curriculum.

3

The unreliability of grades as measures of academic ability, due to their use for disciplinary purposes, their indefinite meaning, and their serious constant errors, makes their use as a criterion of ability questionable.* We shall, therefore, attempt to find a satisfactory method of obtaining estimations of ability from instructors.

The method of obtaining estimations used by earlier investigators, e.g., Dressler,[6] Gilbert,[9] and Kirkpatrick,[12] was to ask that the student be placed in one of the three classes, *good, medium,* or *poor.* This division may be satisfactory if one wishes only an indication of the differences between the averages of the three classes in some particular performance, but it is valueless in specifying how well or at what point a *division* of the entire group may be made. Such a classification is based upon the assumption that all instructors will conceive identical points dividing the undefined classes *good* and *medium,* and *medium* and *poor.* Although the judgments are easy to make by this method, for our purpose the resulting divisions are entirely too rough.

A second method and a very important one is the *order of merit* or *rank method.* This method has been used in many investigations; among the most important are Cattell's[5] studies of American men of science, and Spearman's[21] and Burt's[3] studies of general intelligence. The judge is usually instructed to select from the group the individual whom he considers foremost in respect to the quality in question; and to place this individual first. From the remainder of the group the best is again chosen, and is placed next in rank. Sometimes the procedure is varied, and the poorest is selected after the best. When great accuracy is desired, and there are not too many individuals, a modification of the *order of merit method,* the *method of paired comparisons,* may be used. By this method every individual is paired with every other one, and in all cases where an individual excels the person with whom he is paired, he is given a plus mark. The

*If homogeneity in performance were the sort of homogeneity desired, grades might be used as the criterion of the value of mental tests.

individuals are finally arranged according to the number of plus marks they receive. The increase in accuracy resulting from this elaboration of the method is probably not great enough to compensate for the added time required to make the judgments. Certainly the number of subjects required for an investigation of this kind precludes the use of the method of paired comparisons.

The order of merit method does not offer many difficulties in situations where it can be used, and it has some noteworthy advantages. The method permits averaging the judgments on any one individual so that the harmful effect of accidental errors may be reduced. Furthermore, the probable difference between the positions of two individuals may be calculated. Both of these advantages of the order of merit method were shown by Cattell.[5] In his research, the ten foremost men in each science were required to place from one hundred to three hundred of their colleagues in the order of their ability. Suppose one individual received the ranks of 10 15 16 12 14 21 15 19 12 11; an average of these ranks may be computed to give his average rank. The probable error of this individual's rank is obtained from the variability of the judgments upon him. "The difference in scientific merit between any two of the psychologists . . is directly as the distance between them, and inversely as their probable errors. If two of them are close together on the scale, and if the probable errors are large, the difference between them is small, and conversely." If the average ranks of all are determined, an order of merit may be arranged on the basis of the average ranks.

In common with grades, the order of merit method gives an exact evaluation of the ability of each individual, and we are not compelled beforehand to accept, in each of our final classes, a definite percentage of the entire group.

An objection often urged against the order of merit method is that two individuals of average ability appear as widely separated as two individuals of extreme ability. This is a valid objection to the method as it is usually used, but there are certain devices whereby the difficulty may be removed. One is Cattell's plan for

estimating the differences between individuals from the probable errors of the judgments upon them. Another, which has been used extensively by Thorndike,[23] is that of measuring the differences between two qualities on a scale in terms of the ratio of the judgments of superiority to judgments of inferiority, on one of the qualities. When the ratio is 3, the difference is called a *unit* difference. A third method was employed by Galton[7] in his study of hereditary genius. If it is assumed that individual abilities distribute in the form of the probability curve, we may arrange our individuals in the probability curve, and determine just what point in the scale corresponds to every person. This has the effect of decreasing the differences between mediocre individuals, and of increasing the differences between individuals at the extremes. The labor involved in making this correction is considerable, but it may be lessened by the use of tables.

In spite of its advantages, the order of merit method has one limitation which has not yet been satisfactorily overcome. It requires that every judge have sufficient information about every individual to justify an estimation. For the method makes it imperative that every student be ranked by every instructor. There may be no partial lists; and the fact that a judge is not acquainted with one man is enough to discard technically the rest of his rankings. In the concrete situation, where the instructor has only 20 or 25 students in a class, his contribution to a final order of merit would be quite valueless. In researches where moderately large groups have been tested, and this difficulty has become acute, it has been met by using the rankings of only one judge, and by neglecting the records of whatever students he failed to estimate. This device loses for the order of merit method many of its advantages, and makes the ranks actually obtained subject to many errors. Unless the limitation here mentioned can be removed, the order of merit method cannot be used in mental test investigations which use judgments as their criterion and attempt to consider even a medium number of subjects.

Thorndike[24] has proposed a method for approximating the true order of merit where the series of judgments are partial. It is impossible to give an exposition of the method without the

reproduction of many tables, and so the reader is referred to the original article. The method, unfortunately, is not applicable to our conditions, for it requires that each individual be estimated by many judges, and during a semester it is unlikely that a student will have more than four or five instructors. It is further necessary that the different instructors judge many individuals in common, and in a freshman class that numbers more than one hundred, the required overlapping may not be found. Thorndike's method is probably applicable in certain circumstances, but it does not make the order of merit method available for securing estimations of ability of the members of a freshman class.

A further disadvantage of the order of merit method is that it does not permit the combination of two separate groups into a single group of large size.

The failure of the order of merit method is due to the fact that the position of each individual is conditioned by the qualities of the other members of the same group. It seems clear that the only way to obtain the judgments is through the use of some scale that is external to and independent of any particular group. To meet this demand there are two kinds of scales: one such as that used by Webb,[26] and another as used by Pearson.[15]

Webb's plan for forming scales is not limited to scales of intelligence, but is applicable to practically any situation where estimations of qualities, e.g., honesty, sense of humor, is desired. The following excerpt from Webb's monograph describes the method.

"The following instructions were issued to all judges:

1. Personal qualities are named and briefly annotated in this schedule. If you have any doubt as to the meaning of any of them, please ask me.

2. In the columns under each subject's name place one of the marks

$+3 \quad +2 \quad +1 \quad 0 \quad -1 \quad -2 \quad -3$

for each of the qualities specified. To avoid errors, please put the + signs as well as the —.

3. The mark +3 is for those showing a very high degree of the quality as compared with the average.

+2 is for those showing a degree of the quality distinctly above the average.

+1 is for those showing a degree of the quality slightly above the average.

0 is for those possessing the average degree of the quality for the group you are judging.

—1 is for those slightly below the average.

—2 is for those distinctly below the average.

—3 is for those showing the lowest degree of the quality as compared with the average.

4. As far as it is possible in your group of 20 men, the number of subjects receiving the above marks should be 1, 2, 4, 6, 4, 2, 1 respectively."

The limitation in the fourth point is an attempt to force all the estimations into the form of the probability curve.

Judgments made on this plan may be averaged, and the seven classes seem to offer a sufficient number of points on the scale. There are however three objections to the scale which cause us to reject it for the purposes of this research.

1. It is not legitimate to force a group of twenty or thirty individuals into any specific form of distribution. Assuming that the total population distributes according to the probability curve, we might still expect a great variety of distribution in samples from the total population which contain only 20 cases. This expectation is revealed by the high probable errors of the frequency constants of distributions containing so few individuals. Since the estimations from any one judge should not greatly exceed 20, the provision contained in the fourth point is not well advised.

2. The scale is not sufficiently objective, since the 0 mark is given to the individuals possessing an "average degree of the quality *for the group you are judging.*" This limitation on the objectivity of the scale would not be serious if all judges estimated precisely the same group of individuals. But in our situation, it is possible that an individual might receive very different marks, not because of any difference of opinion as to his academic ability, but simply because in one case he happened to be

a member of a group of high average ability, and in the second case a member of a group of low average ability.

3. There is an assumption that the differences between the steps of the scale are equal. This may conceivably be true, but the point must be demonstrated. If the steps are not equal, many errors in the position of individuals would creep in, due to the inaccuracy of the averages of the judgments. Webb says, "The most reasonable bases appear to be given by taking the seven classes as equidistant from one another. This has the effect of making the distribution approximately normal." If Webb means that equidistant steps will make *any* distribution approximately normal, he is controverted by many facts; if he means that in his particular research equidistant steps make *his* distribution normal, we need not be too greatly impressed, for he has loaded his dice by asking that the judgments be given *according to the frequency required by the normal curve for equal steps.* Of course, under such circumstances, equidistant steps have the effect of making the distribution approximately normal. The form of the distribution of the judgments is not evidence for equal steps, but is a consequence of assuming them.

For these reasons, the Webb scale is abandoned. There still remains as a possible method of securing estimations of ability the Pearson scale of intelligence.

There are three important related facts that enter into the discussion of any frequency distribution, and that must be considered in the formation of any scale. These are 1. the form (or equation) of the distribution curve; 2. the relative distance between the steps of the scale by which measurement is made; 3. the relative frequency, or the number of individuals, at each point of the scale. For example, Galton[7] assumed that abilities followed the probability curve, and that the steps between certain classes on a scale of abilities were equal; from these assumptions he could tell what the number of individuals in each class should be. In mental measurement, it is a familiar practice to assume that the steps of a scale of measurement are equal (a questionable assumption), to determine the frequency with which observations fall at any point of the scale, and then to draw conclusions concerning the form of the distribution.

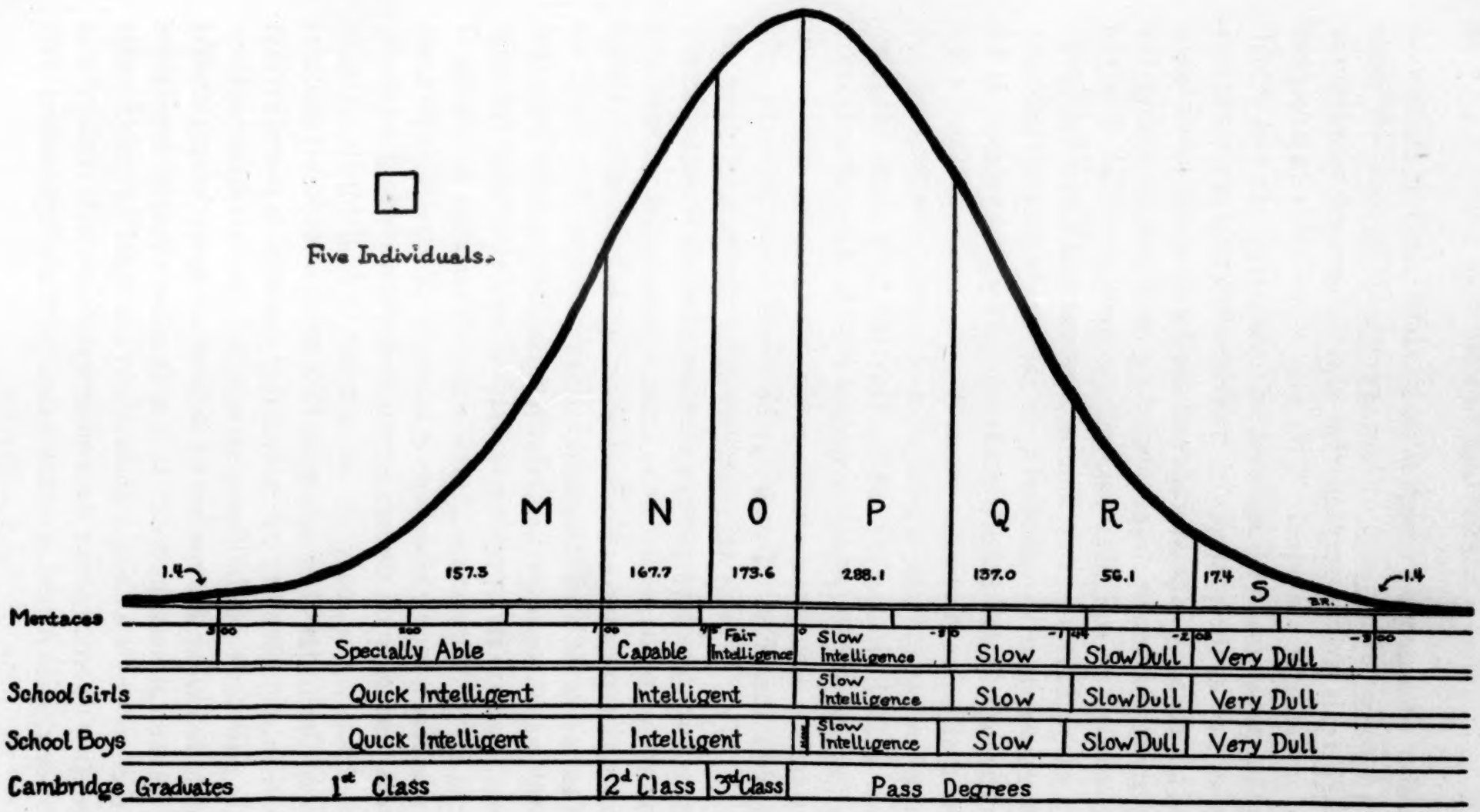

FIG. 1.—The Pearson Scale of General Intelligence. The distribution curve. For convenience, 300 mentaces is usually taken as the mean instead of 0. The standard deviation is 100 mentaces. After Karl Pearson, Biometrika V, 1906, p. 110.

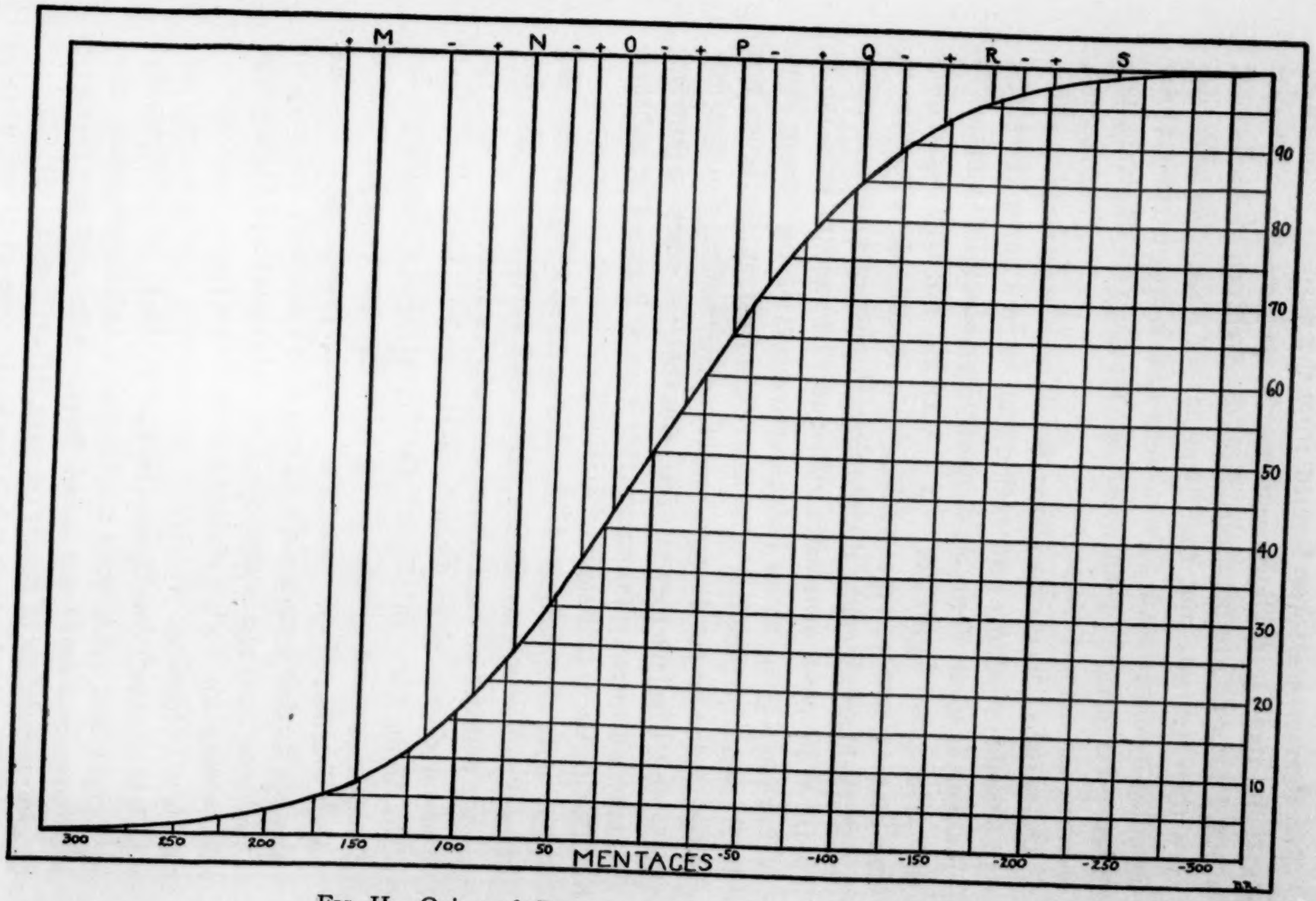

FIG. II.—Ogive of Pearson Scale of General Intelligence.

The Pearson scale was formed from the third of the possible relations: that is, intelligence in general was assumed to be distributed according to the probability curve; classes of intelligence were then defined; and the frequency of each class was determined by experiment. From these data it was possible to determine the relative distance between the steps of the scale, i.e., between the defined classes.

The assumption that intelligence follows the probability curve will probably need less justification than it did a dozen years ago. However, it seems desirable to quote Pearson on this point. "Is this assumption legitimate? It is certainly not true for organs and characters in *all* types of life. But it really does describe in a remarkable manner the distribution of most characters in mankind. We have shown within the limits of random sampling it is true for a great variety of measurements on the human skull . . . I should be the last to assert that no human characters can be found that do not diverge sensibly from the Gaussian distribution. But I believe they are few, and that for practical purposes we may with nearly absolute safety assume it as a first approximation to the actual state of affairs."

The next step was to define classes of intelligence that fall naturally into a quantitative scale, and to determine the frequency with which each class is actually estimated.

Seven classes were defined on the scale: *Very Dull, Slow Dull, Slow, Slow Intelligent, Intelligent, Quick Intelligent* and *Erratic-Inaccurate.* The *Erratic-Inaccurates* turned out to be only a fraction of one percent of the total group and so the class was not further considered. It was found desirable to divide the *Intelligent* class again into the *Fair Intelligent* and the *Capable.* The definitions of the classes are as follows:

Class L *Genius.*

Class M *Quick Intelligent* (Specially Able): A mind especially bright and quick both in perception and in reasoning about not only customary but novel facts. Able and accustomed to reason rightly about things on pure self-initiative.

Class N *Capable:* A mind less likely than M to originate inquiry, but quick in perception and reasoning about the perceived.

Class O *Fair Intelligent:* A mind ready to grasp and capable of perceiving facts in most fields. Capable of good reasoning with moderate effort. This group comprises, say, one third of the total population.

Class P *Slow Intelligent:* A mind slow generally, although possibly more rapid in certain fields, but quite sure of knowledge once acquired.

Class Q *Slow:* A mind advancing in general, but very slowly. With time and considerable effort not incapable of progress. Very slow in thought generally, but with time understanding is reached.

Class R *Slow Dull:* A mind capable of perceiving relationships between facts in some few fields with long and continuous effort, but not generally or without external aid.

Class S *Very Dull:* A mind capable of holding only the simplest facts, and incapable of reasoning about or grasping the relationship between facts. This group passes into the mentally defective.

Class T *Imbecile.*

Estimations were obtained by Pearson on 4638 school children of various ages. Great care was taken to gather the data from all portions of the United Kingdom, so that the sampling might be as free from local influence as possible. The data on the boys were kept separate from those on the girls so that two distinct groups were obtained.

A further group was used for reference which consisted of 1011 Cambridge graduates. These were classed in four grades: *First Class Honours, Second Class Honours, Third Class Honours,* and *Pass Degrees.*

The median point in the judgments on the boys was found to lie approximately between the *Intelligent* and *Slow Intelligent* groups. In the girls, this point was closely between the same two groups, and in the class of the Cambridge graduates it lay between the *Third Class Honours* and the *Pass Degrees.* It was further found that the frequency of the *Intelligent* group was practically equivalent to that of the *Second and Third Class Honours;* and on this basis Pearson suggests that the *Intelligent* class be divided into the two subgroups.

Since there is no reason to suppose that men and women are equally variable in intelligence, the distribution of the three groups in the probability curve was made on the assumption that the common unit of the scale for the three groups was the *Intelligent* class. This class was taken to be a common unit on the scale for the boys and for the girls; and it was equal to the range of the Cambridge graduates who received Second Class and Third Class Honours.*

It was found that when these three groups of individuals,—the boys, the girls, and the Cambridge graduates,—were distributed on the normal curve, according to the plan mentioned above, the agreement of the limits of the defined classes was very close. The three groups were combined into one, and the limits of the classes on the scale were found by determining the deviations from the average, measured in terms of the standard deviation, which were required to include in each class the observed percentage frequency. The numerical value for each class was found by determining the point on the scale corresponding to the mean value of the class.

The final units of the scale were made by dividing the range of the *Intelligent* class into one hundred parts, which are called *mentaces*. Average intelligence was put at 300 mentaces. It was then possible to express the numerical value and the range of each class in mentaces, since the relative ranges of the classes were already known. The standard deviation was found to be 93.3 mentaces. Pearson suggests, however, that 100 mentaces be taken as the standard deviation of intelligence in the formation of his scale, since this round number is a sufficiently close approximation to the true value, and lends itself more easily to calculation.

Two objections to the method of forming this scale must be mentioned. There is some reason for believing that although the form of the distribution of the Cambridge graduates may be the same as that of the children, the central tendency may be different. There must be some elimination of the lower grades of intelligence before British school children reach the university.

*[15] P. 108.

This is not a serious objection, for if the scale were based on the results from the boys and girls alone, it would be practically unchanged.

It is also to be noticed that the division of the *Intelligent* class into subclasses was made after the estimations had been collected, and hence these classes as finally defined were not objectively measured. However, the ranges of the classes on either side of the *Intelligent* class were determined, and so it is probable that the only error would arise in the point of division of the two subclasses. In a later paper slight changes were made in the wording of two of the classes, but the effect of these changes upon the scale is probably inappreciable.

The question now arises whether a scale formed on the basis of estimations of children's intelligence is suitable for the purpose of securing estimations of college freshmen. If there is found no relation between intelligence as judged by the scale and age, then there can be no objection to the use of the scale. The children judged varied in age from 4 to 20, the ages between 6 and 18 each including 39 cases or more. The correlation ratio between age and intelligence was found to be for boys —.054, and for girls—.081. These relations are practically inappreciable, and the ages covered extend well into the period in which we are interested. There seems, therefore, to be no objection on these grounds to the use of the scale.

4

We have not yet examined the errors that may occur in estimations, obtained either by this plan or by any other. The errors that are really serious, the constant errors, are probably few when estimations are used, because a *variable* allowance is made when abilities are estimated for the factors that produce the constant errors in the case of grades, and errors from these sources may therefore be considered of the accidental type. If a student were very modest, and kept rigidly to a speechless part in all his work, he would probably be underestimated; and yet, a bright student of this kind would do so surprisingly well in written work that the harmful effect of his attitude might con-

ceivably be removed after a semester of study. The opposite type, the affable, diplomatic student, might be overestimated; yet the question might be raised whether a student who is able to fool all of his instructors by sheer affability and diplomacy is really so very stupid. Cases of *universal* overestimation or underestimation of a student would probably be very rare. And only in so far as the errors are in the *same* direction may they be called *constant* errors.

One important source of constant errors in the use of estimations comes through the conversations of instructors concerning students. It is probable that in a few cases, a set of judgments on an individual might be considerably in error due to a strong and expressed favorable or unfavorable reaction on the part of one of his instructors. Errors of this sort will be more frequent in a small school than in a large one; they will also be more frequent in judgments on juniors and seniors than in estimations of freshmen.

But if the constant errors are few, the accidental errors may conceivably be very, very numerous. Let us see how many sources of accidental error we can recognize. 1. Errors might arise in case different judges interpreted the definitions of the classes differently. 2. The definitions of the classes might be ignored, and the judges might evaluate the abilities of the individuals according to their own idea of intelligence. 3. A judge might feel that he knew a student well enough to form an estimate, when, as a matter of fact, the student's real ability had never been shown. 4. Judges might make too great or too little allowance for the student's social interests, his athletic ability, his nervousness in recitation, his absences. 5. There is opportunity for all manner of personal prejudice, both favorable and unfavorable. 6. A student with exceptional interest in one branch might be judged too high by one instructor and too low by every other one. 7. Besides all these sources of error, there are the countless small mistakes that are bound to occur in every situation where an estimation is required. To be sure, these are technically *accidental errors;* yet if their number and importance were too great we should have little faith in the mathematical machinery devised to cope with them.

There is an extremely simple and definite method of estimating the importance of the accidental errors. This is the observation of the size of the average of the variations of the judgments on the students. For although compensatory errors have the effect of making the difference between the average of the estimations and the true value approximately zero, *they cannot operate to make the variability of the estimations vanish. On the contrary, the variability may be increased.* A simple illustration of this is found in the following condition. Suppose an object to have a true value, 8; and suppose two measurements to be made on the object, one with the result of 8, and another with the result of 12. The average of the measurements is 10 and the average deviation or variability is 2. If a third measure which contains a compensatory error, say 4, be made, the average becomes 8, which is the true value; but the variability is increased from 2 to 3 2-3. Thus accidental errors neutralize one another in the average, but they cannot do this in the average deviation. The only accidental errors that will reduce the average deviation are those which cause a measurement very close to the average of the previous measures. If these previous measures have been influenced principally by accidental errors, their average will be near the true value, and the new measure which is near this average can consequently contain *only a small accidental error.* Thus it follows that the size of the average deviation denotes the magnitude of the accidental errors which have affected the actual estimations.

To determine how serious the accidental errors are (we grant readily that there may be many of them) we have computed the average of the variability of the judgments on an individual for two groups of college freshmen. Since some variability might be expected, due to the fact that the judges have observed the students under different conditions, it is astonishing to find that the average variability of judgments in one group containing 39 students was only three-fifths of the average difference between the major classes of the Pearson scale. In a second group of 52 students the average variability of judgments was only two-fifths

of the average difference between the classes.[1] The size of this variability may be still better judged by noting that if three judges estimated each person, and that if one judge had a tendency to estimate one class too high, *even though there were no other errors,* the variability from this cause alone would be four-ninths of the average difference between the classes. These variabilities are similar to those found by Waite who used the same scale. From Waite's[25] data, comprising 3427 pairs of estimations, the average variability seems to have been about three-fifths of a class interval. From this experimental work it is possible to conclude that the unreliability of the judgments due to accidental errors is probably not great, certainly not great enough to discourage the use of the scale.

During the preliminary experimental work of adapting the Pearson scale to the purposes of this research, it was at once evident that the seven classes provided by Pearson did not permit as much differentiation as the judges were able to make. Accordingly, the plan was adopted of allowing the instructor to add to the letter designating any class + or —, if the student seemed to him to belong in one class with a marked leaning in either one direction or the other. The majority of the judges were satisfied to use the straight symbols, yet the + and the — were used often enough to make their retention as part of the scale advisable. The evaluation of the + and — classes is not difficult. The mean of each major class is found, and the number of cases falling between each pair of adjacent averages is found from the probability integral. Each group of cases is then divided into three parts, and the two points of division on the scale give the numerical value of the + and — subclasses.

With the scale in this form, the following directions for making the estimations of ability were sent to instructors:

"In order to standardize judgments on mental ability, the following classification of intelligence has been selected. Please note that the scale covers the range of the population at large from the genius to the imbecile.

*Judgments on a further group indicate that two-fifths of a class interval is a lower value than will ordinarily be found. Three-fifths is more to be expected.

On the accompanying blanks, please place the letter standing for the class in which you judge the student's mental capacities to fall. What is desired is a judgment of general intelligence and not of classroom performance. If the individual seems to be in one class with a marked leaning toward another class, judge him to be in the more certain class, and indicate the direction of his leaning by + or —."

The definitions of the classes of intelligence followed. Letters M, N, O, etc., were used to indicate the various classes which could not be easily confused with the letters used for grades in the University of Chicago, i.e., A, B, C, D, and E.

No judgments were made until the conclusion of the term, in many cases of the school year. It is encouraging to note that several instructors returned blanks with the remark that they felt unable to make judgments on some students due to prejudice or lack of information.

As soon as the estimations were received the letters were translated into their numerical equivalents, and the average of the judgments and the average variation of judgments were found for each student.

The translation is made by the following table:

Table I

M+471	mentaces	N—353	mentaces	P 262	mentaces	R+157	mentaces
M 451	"	O+337	"	P—242	"	R 130	"
M—416	"	O 322	"	Q+220	"	R—116	"
N+391	"	O—302	"	Q 192	"	S+ 96	"
N 371	"	P+282	"	Q—177	"	S 62	"

The average of the judgments on an individual is the measure of his ability that is finally used.

The following is a summary of the standings of two groups. Group J_1 consists of 52 freshmen; group J_2 consists of 39 freshmen.

Table II

Groups	Mean Standing	Median Standing	Standard Deviation	Range
J_1	321.2	322	45.1	439-196
J_2	331.8	331	44.1	451-221

These results show striking likenesses. The students were

taken from two separate institutions and different standards of judging might have been expected to have influenced the average and the dispersion. The similarities are strong evidence for the objectivity of the Pearson scale.

Comparing these data with the general scale, we find that the average has been increased by about 25 mentaces. This increase is probably not so large as might have been expected. There is scarcely any tendency toward skewness, as is shown by the closeness of the median and the mean. Even the slight skewness is found to be in the opposite direction in the two cases. The standard deviation is only about half that of the population at large. A reduction in dispersion would be expected, for the extreme cases at the lower end do not occur, and the range is consequently considerably less than is that of the entire population.

The amount of objectivity which the Pearson scale seems to show might appear at first glance astonishingly great. The actual verbal definitions of the classes of intelligence seem on analysis to be susceptible to a variety of interpretations, and it might be questioned on a priori grounds whether the scale offers a method of subdividing the conceptual range of intelligence in a manner that would be uniform and definite for the different judges. The definiteness probably arises from the repeated association of the terminology of the scale with typical manifestations of intelligence, and from the rather uncritical acceptance of this terminology as an adequate identifying mark of these manifestations. The uniformity of the division of the conceptual range for different judges may come from the similarity of the environments in which the terminology acquired its meaning for the different judges. Had the judges been drawn from all walks of life, it is probable that less uniformity would have accompanied the use of the scale.

On the whole the scale gives results which seem satisfactory. We may therefore with good reason accept the Pearson scale as the best available method of obtaining judgments of the general ability of students; and we may use the judgments thus obtained as the criterion of the value of mental tests in sorting from a mixed freshman class smaller groups of greater general intellectual homogeneity.

5

It may be well to add at this point that such judgments may be put to other uses than that of testing the value of the tests. The judgments should be continued through the college course, and changes from term to term in the judgments on any student should be watched. Waite[25] found that in the majority of cases the change is slight, yet the exceptions will provide material for interesting and profitable study.

Single erratic judgments on students should be strictly observed. If one judge estimates an individual much higher than do the other judges, the possibility of special ability should be investigated. If one instructor estimates an individual far below the judgments of others there is an indication of an unfavorable personal reaction, and if possible, the student should be placed under another instructor.

These applications of the judgments promise that in practical work many other uses for them may be found. The judgments give information that is just as essential in properly advising the freshman as that given by the tests, and since the judgments are so easily obtained, they should form an important part of every student's record.

III

1

Although we have found a criterion of academic ability that will be satisfactory for our purpose, we must still postpone the discussion of the value of the mental tests in selecting groups of students of intellectual homogeneity.

It so happens that at present two researches must be carried on at the same time. First, we must try out many tests, and try to choose those that are the most promising. Second, we must discover how well the chosen tests *as a group of tests* will serve in the diagnosis of ability. It would be better if these two investigations could be carried on separately, but there are obvious difficulties. The evaluation of a *group* of tests is not warranted, for there is no group of tests known that would

justify an unchanged program through a period of years. The evaluation of *single* tests is impracticable, for until psychological tests are shown to have some specific administrative value, the labor involved in applying them in an institution makes their introduction for purely experimental purposes unjustified. Consequently we must point our attention in both directions at once, and if we can find indications that the tests may be used for diagnostic purposes, we may feel certain that there will be opportunity for extended research on the single tests.*

Clearly our first task is to choose groups of tests that seem most likely to give good results. This section will therefore be devoted to the selection of the groups of tests that will finally be evaluated. We do not overlook the problem of evaluating the single tests. But the methods of determining the relative weights of the single tests require that we have *accurate inter-test* correlations; and to secure these extended investigation is necessary.

So long as the tests are still in a period of development, both in respect to their content and their administration, it seems reasonable to confine research to classes of small size. Under these conditions, if a test prove worthless or a method faulty, there will be no serious loss in its discard. And yet at the same time much positive information may be obtained, even with small groups. To be sure it will be impossible to say that our results, particularly those concerning the points at which the group may be best divided, are valid for freshmen classes containing hundreds of members. But the converse is also true; for mental tests which are very satisfactory for large groups might give quite erroneous information if they were used in exactly the same way for a small and more rigidly selected group.

Two freshmen classes were tested which, though small, yielded a number of cases well above the minimum required for correlation. As will be seen from the description of the groups which follows, they are alike in little except size, and therefore, such results as are in harmony, may be taken as suggestive of what may be expected from mental tests when they are used on small freshman classes.

*[22] Ch. V.

The two groups will hereafter be called J_1 and J_2. (The letter J is chosen to distinguish these two groups for which judgments were obtained, from a group G_1, for which grades were used as the criterion of ability.) J_1 is a first year class of the School of Commerce and Administration at the University of Chicago. The entrance requirement is high,—15 units and no conditions from an accredited high school, with an average grade of 80. Since the school gives a particular form of professional training, it is probable that the interests of the individuals are very much the same. J_1 contains 52 freshmen; of these 40 are men. The average number of judgments on an individual is 3.06, with a minimum of 2.

The second group of freshmen, J_2, is taken from a middle western women's college. (The name of the institution is withheld by request.) J_2 contains 39 individuals. The entrance requirements for admission to this college are about the same as those for admission to the School of Commerce and Administration,—graduation from an accredited high school. Admission is still further restricted by an entrance examination. There are many differences between J_1 and J_2. J_2 is made up entirely of women, while J_1 is for the most part composed of men. The majority of the members of J_2 have their homes in the town in which the college is located, and the group is therefore probably more homogeneous with respect to previous experience than is J_1. The average number of estimations on the members of J_2 is 3.69, with a minimum of 2. In all, 70 students of this college were judged in order to procure information about the Pearson scale.

A third group, hereafter called G_1, consists of Commerce and Administration freshmen. This group was tested by Dr. H. D. Kitson,[13] and the results of this testing are the basis of his *Scientific Study of the College Student*. To him we are indebted for the use of the data. G_1 contains 50 freshmen; of these 41 are men. Unfortunately at the time this research was begun the group was so broken that it was impossible to secure estimations of the students' abilities, and so grades were used as the criterion. G_1 will be used principally for purposes of comparison with J_1 and J_2.

There is an important difference between J_1 and G_1 that must be recognized. The attitude of J_1 toward the academic work was much more serious than was that of G_1. The former group was given a three weeks' course in methods and ideals of study; extra-curricular activities were discouraged, and every effort was made to raise the standard of the classroom work. When we examine the relation of grades to the mental tests, this difference in attitude will be important in our interpretations.

The following tests were used in one or another of the groups.

Hard Directions	Logical Memory Visual
Absurdities	" " " Deferred
Sentences Built	*Constant Increment
Opposites A	*Cancellation
Opposites B	*Business Ingenuity
Analogies A	*Words Built
Analogies B	*Numbers Heard
Alphabet	*Objects Seen
Logical Memory Auditory	*Oral Instructions
Logical Memory Auditory Deferred	

The tests marked * are described in Kitson's *Scientific Study of the College Student.*[13] The remainder of the tests are described in Appendix I.

Table III gives the means and standard deviations of the tests for groups J_1 and J_2. The probable errors are too high to warrant a discussion of differences between the groups.

Table IV gives the product-moment coefficient of correlation between each test and the criterion of academic ability used for the group.*

In spite of the fact that there is considerable variability throughout Table IV, the first group includes the Hard Directions, the Absurdities, the Sentences Built, the Alphabet, the Opposites, the Analogies, the Logical Memory, and the Constant

*The student's standing according to grade in courses is the average number of grade points obtained during the academic year. A grade of A gives 6 points, A— 5 points, B 4 points, B— 3 points, C 2 points, C— 1 point, D 0 points, E— 1 point, and F— 2 points. The total number of grade points of each student was divided by the number of academic units (or majors) that this student would obtain by passing all courses pursued by him during the year.

Increment tests. These tests are unquestionably the best of the nineteen for diagnostic purposes judged on the basis of their correlations with estimated ability. There is a second group of tests that appears to be extremely erratic. The tests of this group are the Cancellation and the Business Ingenuity tests, and it seems that these tests must be further investigated before they can be put on a par with the tests in the first group. In the third group are the Words Built, the Numbers Heard, the Objects Seen, and the Oral Instructions. These tests are relatively inferior.

TABLE III

Test			Group	Mean	Standard Deviation
Hard Directions	Time		J_1	115.9 sec.	31.8 sec.
			J_2	124.8	38.5
	Errors		J_1	2.54 errors	2.08 errors
			J_2	1.76	1.76
Absurdities°	Time		J_1	87.3 sec.	26.2 sec.
			J_2	103.3 sec.	38.8
	Errors		J_1	3.42 errors	1.50 errors
			J_2	2.84	1.13
Sentences Built			J_2	6.7 sentcs.	2.09 sentcs.
Opposites	A	Time	J_1	26.5 sec.	6.87 sec.
			J_2	25.4	6.25
	B		J_2	43.1	13.3
Analogies	A	Time	J_2	69.8	26.2
	B		J_2	69.7	24.1
Alphabet			J_1	70.7	14.7
			J_2	69.5	18.9
Logical Memory°			J_1		
Auditory				45.5 points	17.5 points
Auditory deferred				28.6	21.8
Visual				62.5	16.1
Visual deferred				41.0	16.1
Logical Memory			J_2		
Auditory				79.2	13.6
Visual				85.9	13.5
Constant Increment			J_2	160.9 sec.	35.9 sec.
Number Checking			J_1	64.3 checks	11.8 checks
			J_2	58.2	10.1
Business Ingenuity			J_1	28.2 points	11.7 points
Words Built			J_1	19.6 words	3.78 words
			J_2	20.3	4.19
Numbers Heard			J_1	8.82 digits	1.90 digits
Objects Seen			J_1	7.36 objects	1.10 objects

° In this test a part of the differences in the means and standard deviations of the two groups may be attributed to differences in the test. See Appendix I.

TABLE IV

Correlation between Standing in Tests and Criteria of Academic Ability.

	G_1		J_1		J_2	
	r	P.E.	r	P.E.	r	P.E.
Hard Directions Index	+.38	.07	+.54	.07	+.39	.09
Absurdities Index			+.52	.07	+.47	.08
Sentences Built	+.27	.09			+.43	.09
Opposites A	+.24	.09	+.38	.08	+.20	.10
Opposites B					+.37	.09
Analogies A					+.44	.09
Analogies B					+.36	.09
Alphabet			+.50	.07	+.18	.10
Auditory Memory	+.37	.08	+.35	.08	+.16	.11
Aud. Mem. Deferred	+.31	.08	+.45	.07		
Visual Memory	+.20	.09	+.26	.09	+.47	.08
Vis. Mem. Deferred	+.26	.09	+.29	.09		
Constant Increment	+.28	.09			+.29	.10
Number Checking	+.23	.09	+.26	.09	—.20	.10
Business Ingenuity	+.03	.09	+.36	.08		
Words Built	+.06	.09	+.11	.09	+.05	.11
Numbers Heard	+.09	.09	+.13	.09		
Objects Seen	—.08	.09	+.09	.09		
Oral Instructions	+.06	.09				
Standing in Tests Combined	+.43	.07	+.65	.06	+67	.06

No discussion of differences between the tests included in any one of these three main groups is justified because of the high probable errors of the correlation coefficients.

We are however able to choose the tests that may best be included in the groups of tests whose worth we shall finally determine.

In the series of tests which will be evaluated with respect to the estimated abilities of J_1 are included: Hard Directions, Absurdities, Alphabet, Opposites A, Logical Memory Auditory, Logical Memory Auditory Deferred, Logical Memory Visual, Logical Memory Visual Deferred. The group of tests will be called Test Series J_1.

In the series of tests which will be evaluated with respect to the estimated abilities of J_2 are included: Hard Directions, Absurdities, Sentences Built, Opposites A and B, Analogies A and B, Alphabet, Logical Memory Auditory, and Logical Memory Visual. This group of tests will be called Test Series J_2.

For the G_1 group, all the tests which were given were included in the series which was evaluated with respect to grades.

2

There are many problems concerning the actual administration of the tests that are worthy of comment. We have seen that the greatest value of mental tests comes in the *immediate* information that can be gained from them. Then, too, the rapidity with which a knowledge of the tests spreads through a college body makes it imperative that the testing be extended over the shortest possible time. These demands for speed in the application of the tests make the choice of the tests and of the conditions under which the tests are given a very important matter.

To guard against the spread of information about the tests, the psychologist should speak to the freshmen, concerning the value of accurate mental examinations. An appeal should be made to the class for its cooperation. The effectiveness of the appeal will depend partly upon the manner in which it is made, and partly upon the length of time which is required for the testing of the class. Each student should be questioned when he is tested concerning his knowledge of the tests.

The method of testing must be modified according to the size of the group that is to be tested. For classes of medium size, it is desirable to devote two periods to group tests, in order that the Deferred Logical Memory tests may be given. It is also desirable to secure measurements on the students at different times, so that disturbances from temporary indisposition may be lessened. The tests that are best given to the group as a whole are the Logical Memory tests, and the Sentences Built test. The tests that are most suitable to individual testing are the Hard Directions, the Absurdities, the Opposites, the Analogies, the Alphabet, and the Constant Increment tests. With clerical assistance, these tests may be given to an individual in about 20 minutes. The individual testing may thus be completed during the week or ten days that elapses between the first and second group tests. The data can be worked into form in three or four days; of this time, the greater part will be spent in the scoring of

the logical memory tests. It is possible in this way to secure the information from the tests within ten days or two weeks after the opening of college.

It may be that conditions will make impossible two group testings. Under such circumstances, the Deferred Logical Memory tests must be omitted. It will be possible to complete the test work in a shorter period of time if these tests are not given, and there is reason to question whether the tests for deferred memory are important enough to compensate for the delay they cause.

For very large groups consisting of from 300 to 1000 freshmen, it will be necessary to have many assistants to do the individual testing or else all the examination must be given to the class as a group. Most of the better tests are not readily adaptable for group work, and so it seems inadvisable to abandon the individual examinations. If tests are to be given in the fall, assistants may be adequately trained in a course in Mental Tests the previous spring. If assistants are employed, the group testing may be given up if it seems desirable to do so. Lack of facilities would make it necessary for the group testing to be given in parts, and there might be some difficulty in bringing enough pressure to bear upon the freshmen so that absences might be few. The student does not feel his responsibility half so keenly when examinations are made by the group plan.

The tests which might advantageously be given to large groups by individual examination are the Hard Directions, the Absurdities, the Sentences Built, the Opposites, the Analogies, the Alphabet, the Constant Increment, and the Logical Memory Visual. Since the subject does not require the attention of the experimenter in the last named test, it is possible to complete the examination in about thirty minutes. A class of 400 could be tested with the aid of ten assistants in four days; if each assistant has been trained to work up the data on the students he has tested, complete information on the entire class should be available in one week after the beginning of the fall semester. This is soon enough after the opening of college to allow for the separation of the freshmen into homogeneous classes for the work of the first semester.

The scoring of the individuals according to their performances in the tests presents as important a problem as does the actual administration of the examinations. A single numerical value for each individual's ability in all the tests combined is obtained by expressing his performance in each test as a deviation from the mean performance in the test, and by expressing the deviation in terms of the standard deviation as a unit. This gives the individual's *standing* in each test, and the sum of these *standings* is the individual's *standing* in all the tests combined. Woodworth has described this method of reducing scores to standings in detail. The labor of reducing the actual scores to standings is not excessive. The standard deviations may be easily found by the formula

$$\text{Standard Deviation} = \sqrt{\frac{\Sigma(\text{scores})^2}{n} - \text{mean}^2}$$

If an adding machine is used for making computations, the standard deviation may be found at the same time as the mean by printing the square of each score in a second column. The deviations may be quickly divided by the standard deviations by the use of Crelle's Calculating Tables. Of course, *if enough individuals are measured, so that grouping in a frequency table is justified, the above formula is no longer a time saving device.*

If assistants are employed in giving the tests, each assistant may report the sum of the scores and the sum of the scores squared for each test on the individuals whom he has tested himself. This information with a knowledge of the total number of students tested will make the computation of all the standard deviations the work of twenty minutes or half an hour. Tables giving the standings for each score may then be given to the assistants who can record the standings on the record cards in a very brief time.

Only individuals with complete records should be included in means and standard deviations upon which standings are based. Otherwise the great advantage of standings in computing coefficients of correlation will be lost. There will be little error in basing the standing of an individual who has not taken all the tests upon the norms of the rest of the group.

After mental tests have been used in an institution for several years, a tentative evaluation of the relative standing of the freshmen may be made almost immediately after the tests have been given, by expressing the scores in terms of the means and standard deviations of the previous years. Tables may be constructed giving the standing corresponding to any score in any test, and the clerk who records the score may record this tentative standing at the same time. The exact standing could be computed later, but it is not likely that any great variations in the relative positions of the freshmen would be found. The preliminary standings would be satisfactory for the division of the students into groups, although they could not be used easily for purposes of correlation.

IV

I

Now that we have found a method whereby the academic ability of a student may be estimated, and have selected two series of tests which, judged by the correlation between standings in a test and estimated academic ability, are relatively superior, we may proceed to the crucial question of this research: *How accurately would the tests have divided the freshman classes into groups of homogeneous academic ability had they actually been in use?*

We wish to know how a percentage—any percentage—of the individuals, representing the superior or inferior extreme of the entire class according to their performance in the series of mental tests, stands in academic ability as indicated by estimations of intelligence. More concretely, to what extent is the 15 per cent of the class which stands highest according to tests also rated highest in ability as judged? The highest 25 percent; or the lowest 15 per cent? Accuracy or inaccuracy of the internal arrangement of these groups is equally acceptable; it is only important that approximately the same individuals be chosen by the tests as are chosen by the judges.

To answer the question of this research, to determine how well

the series of tests would have done the work of separation, and for what points of division of the group the tests would have done the work most accurately, is to measure the changing relation between the group selected by the tests and judgments of academic ability, as the percentage of the total group included in the selected group is changed.

This situation presents an unusual problem in correlation. We demand a statement of the relation between a measured variable (the estimations of ability), and a second variable (the standing in the tests) which is divided at some point into two alternative categories. We should have an inadequate expression of the relation if we treated both variables as *continuous* variables, and computed an index of relationship by the product-moment correlation method, the method of rank differences, or the foot-rule. For to correlate estimations of academic ability with an exact evaluation of the performance of every individual in the test series is to measure the accuracy of the test series in terms of a problem which the test series will never be called upon to solve—namely, the determination of the precise ability of every individual. The purpose of the series of tests is fulfilled if it succeeds merely in picking out the individuals who are superior,—*it need not distinguish between superior individuals.*

2

In order to indicate the closeness of the relation in which we are interested, the relation between a *continuous* variable and a variable divided at some point into *alternative categories,* we have derived (Appendix II) a coefficient which we shall call the *rank-tangential coefficient,* and which we shall designate by the symbol *t*.

$$t = \frac{M(N+1) - 2\,S(R_x)}{M(N-M)}$$

where N is the total number of individuals; M the number of individuals in the selected group; and $S(R_x)$ is the sum of the ranks in variable X (the continuous variable) of the M best or worst individuals according to variable Y.

The rank-tangential coefficient proves to be a good index of relationship in several important respects:

(1) Its meaning is definite, readily understood, and adapted to the concrete situation in which t is to be used.

(2) It varies between +1 and —1, taking these values only when the M best individuals in Y are associated with the M best or worst individuals in X.

(3) Its value is zero if the two variables are independent.

(4) If M = N, t is indeterminate as it logically should be.

(5) The rank-tangential coefficient does not measure the relation between X and Y in terms radically different from those of the common correlation coefficient. When r is low, t will be low on the average; when t is high on the average, r will be high. In chart I, values of t are plotted for different values of M. In J_2, the coefficient of correlation between X and Y is +.67; in J_1, r is +.65; in G_1, r is +.43. In each of these three curves, the values of t are seen to vary around the value of r for that particular group.

(6) Finally, the rank-tangential coefficient is computed with great ease. Examples are given in Appendix II.

It must be emphasized that the relations measured by t and by r are different, and no direct comparisons of the two coefficients are possible, except in certain specific circumstances.

3

The formula may be applied in either of two ways; for the selected group—the variable given by alternative categories—may be taken either as standings according to tests or standings according to estimated academic ability. The coefficient at any point of division may be very different according to the way in which the formula is applied, and the interpretation of the relationship depends upon which variable, test standings or estimated abilities, forms the basis for selecting the group. If the best 25 percent in *tests* be taken, the relation found by this formula between the 25 percent group and the estimated abilities of the students tells how closely the best 25 percent *according to tests* corresponds with the continuous variable, the estimated abilities. If the 25 percent judged best in ability be taken as the selected group, the coefficient tells how closely the best 25 percent in ability corresponds to the continuous variable, now the test standings. These distinctions may seem hardly worth mentioning, but they are of the greatest importance in the interpretation of the relationships. *For if it is desired to know how well the tests would have picked out a good or a poor group, the standings in*

tests must be taken as the variable given by alternative categories. **The opposite method gives information, but it does not tell how well the tests may be expected to work.**

The alternative categories may be formed by dividing the group at any point according to performance in tests, and thus a series of values may be found for the changing degrees of the relationship as a greater and greater percentage of the entire class is included in the good or poor group. The values of the rank-tangential coefficient may then be plotted for 5, 10, 15, etc. percentages taken in the good or the poor group. See charts. From the values of t that are found, we may draw conclusions concerning the efficiency of the mental tests in separating good students or poor students from the total group. We may also discover at what points the group should be divided in order that the tests may do their work most efficiently.

4

Although we have a method of finding the degree of resemblance between any percentage as selected by tests and the judged abilities, we have yet to decide how close a relationship is necessary before the series of tests may be said to have a practical value. A higher or lower relation would be demanded, depending upon the inflexibility or flexibility of the system of division. For example, if it is desired to exclude students from college on the basis of tests, a much higher correlation would be demanded than if the students are to be temporarily classified for the administrator's information according to their academic abilities. As a matter of fact, the worth of a series of mental tests for the purpose of selecting homogeneous groups depends ultimately upon the degree of relation between standing in tests and ability that is thought to be necessary before a division of the class is justified.

In the report of the use of psychological tests at Reed College, Rowland and Lowden[17] remark concerning a correlation of +.37, "There seems to be little doubt that the revised list of tests did make a selection of the better students in Reed College." It seems probable however that a correlation of only

+.37 would be too low to show a striking value for mental tests in selecting a homogeneous group of students.

There is an indirect way by which we may determine what degree of *correlation* would be necessary. We may arbitrarily say that the root mean square error in estimating abilities from test standings shall not exceed a given amount. We may then compute the correlation coefficient which will permit this error by the formula*

$$r = \frac{S_x^2 - S_r^2}{S_x^2}$$

S_r = root mean square error
S_x = standard deviation of est. abilities.

The coefficient thus obtained will be just as arbitrary as one directly chosen. But it seems reasonable to approach the choice of a critical coefficient from the point of view of the allowable error of diagnosis rather than from the more abstract mathematical relationship.

Let us take one half a class interval of the Pearson scale as a permissible root mean square error of estimation. This is about 31 mentaces. Solving the equation for r ($S_x = 43$) we obtain a value +.693. This is the product-moment coefficient of correlation that must be found between standings in tests and estimated abilities in order to justify the use of mental tests in estimating academic ability.

We shall take a rank-tangential coefficient of 70 as the minimum which would justify a division of a group at any point. This value is taken simply because 70 is suggested for the product-moment coefficient, not because there is any definite relation between the product-moment coefficient and the rank-tangential. As remarked above, *any such minimum value is highly arbitrary, depending in large measure upon the use that is to be made of the test information.*

5

The changing values of the rank-tangential coefficient were plotted in the manner described above. As a result the curves shown on charts I and II were obtained. Along the vertical axis are plotted the values of the rank-tangential coefficient; along

* Yule [80] p. 177.

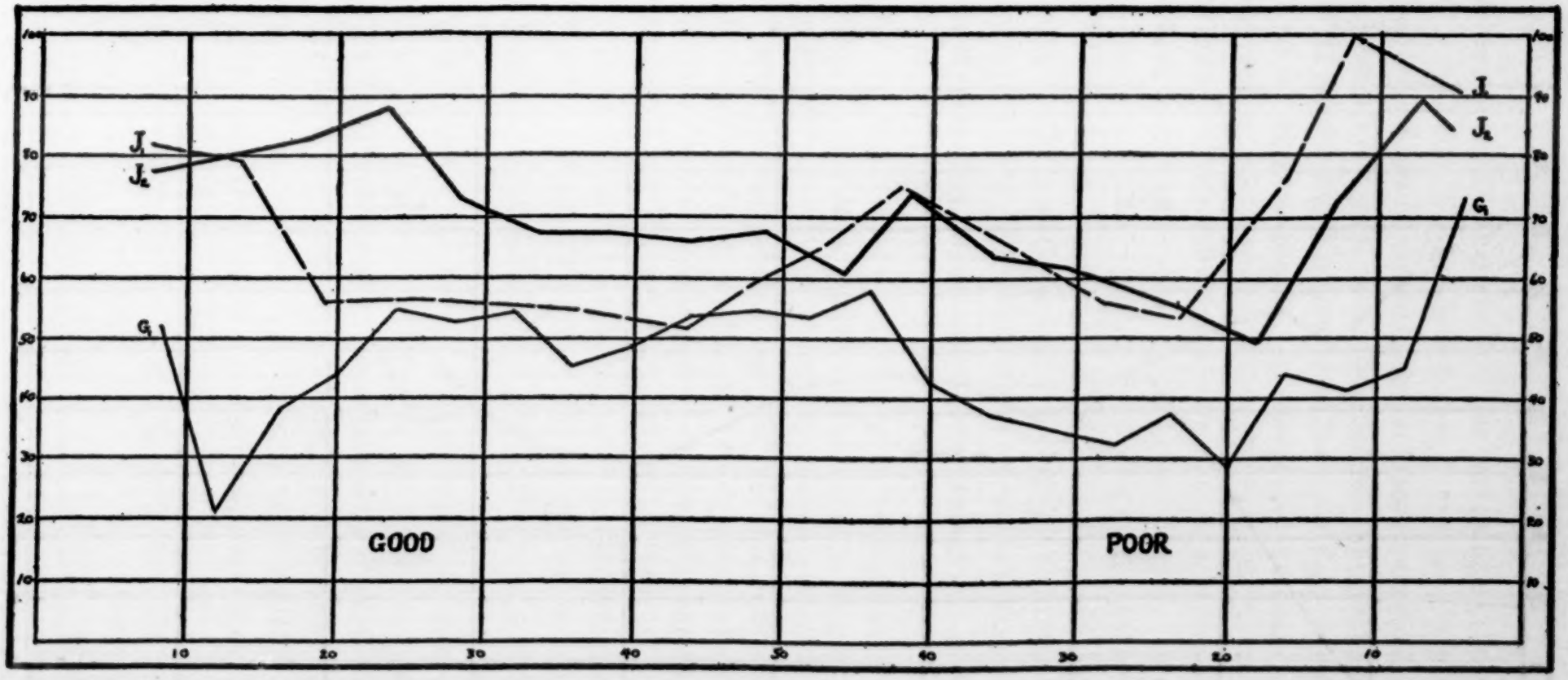

CHART I.—The relation between the rank-tangential coefficient and the percentage of individuals included in the selected group. The three curves represent three different freshman groups described in the text. The ordinates indicate the magnitude of the rank-tangential coefficient; the abscissae the percentage points of division. In each of these curves, standing in tests was taken as the variable given by alternative categories.

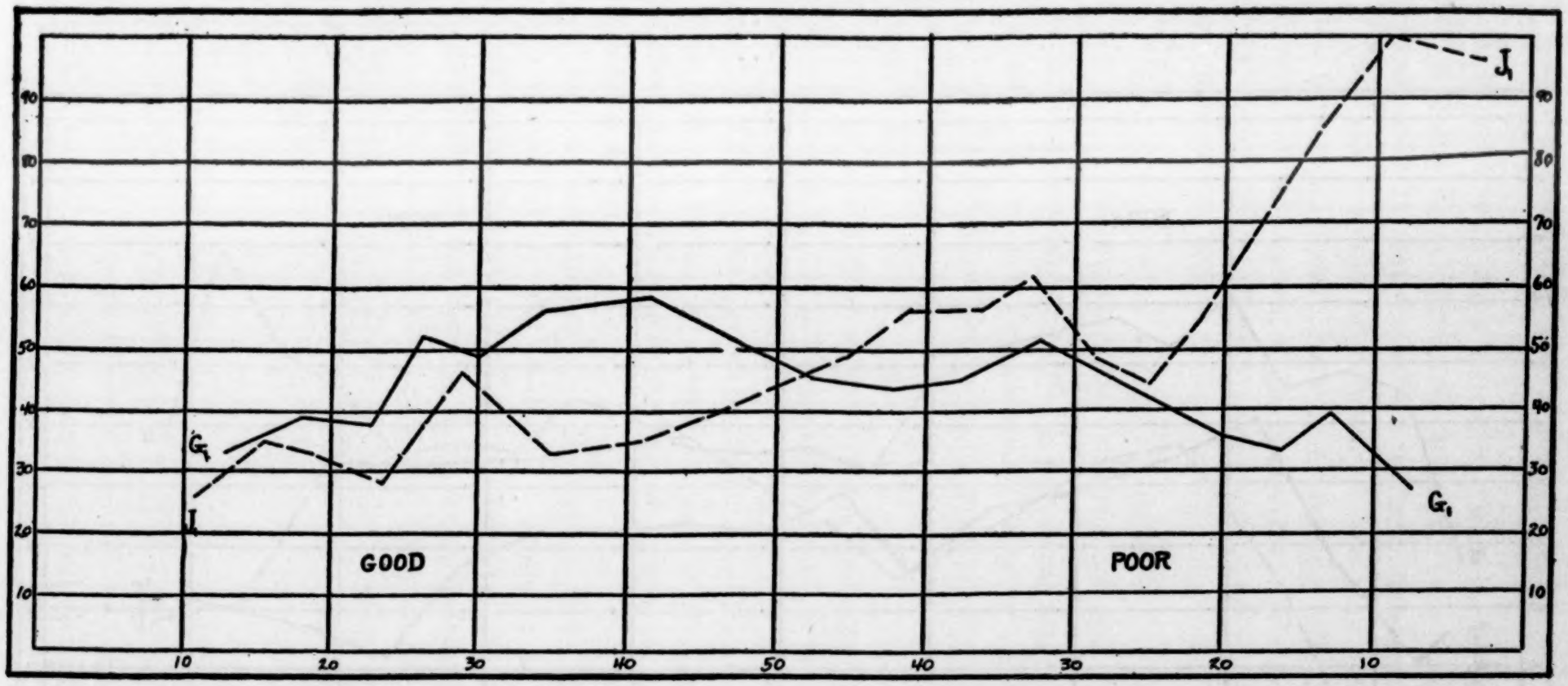

CHART II. In these curves, average grade in courses was taken as the variable given by alternative categories.

the horizontal axis are plotted the percentage points of division. The percentage is always measured from the extreme of the distribution. Thus in Chart I, 20 per cent from the left means that the best 20 per cent were included in the good group; 20 per cent from the right means that the worst 20 per cent were included in the poor group. It was found impossible to make the divisions at any regular intervals, because it often happened that two or three individuals were tied on some value, and so the division would have to be carried over the tie. Fortunately, the ties never extended over an excessive amount of the range.

Chart I shows the values of the rank-tangential coefficient for groups J_1, J_2 and G_1, when the group is divided according to standing in tests. Of these three curves, *the two J curves alone give trustworthy information concerning the value of mental tests in the selection of homogeneous groups.* The G curve is based on the relationship between performance in tests and grades, a criterion known to be subject to frequent and serious errors.

It must be remembered that the probable errors for t are very high (Appendix II), and hence only general similarities of the curves may be commented upon. Slight fluctuations in the value of t in one curve alone are meaningless.

In general it may be said that it is possible to distinguish four groups by the use of tests; a lowest group to include 10 to 15 per cent of the individuals; a poor group to be separated at the lower 40 per cent point, a mediocre group to include the individuals between the lower 40 and upper 15 percentage points; and the individuals who stand in the upper 15 per cent according to the tests. This is seen from the J curves, chart I.

The lower 10-15 per cent of individuals when ranked according to tests are definitely of low degree of ability. The evidence for this is especially good. Not only are the rank-tangentials for both curves high, but they reach this maximum at practically the same percentage point. It is fortunate that the lowest division includes such a small percentage of the entire freshman class. This extreme low section as tested evidently consists of those who should not be in college; it includes the majority of those

who are forced out of college at the conclusion of the first term. If we could be sure of as high a rank-tangential as that shown by J_1, we could safely exclude the group automatically, before it begins its disastrous college career. The correlation will probably lie between the points of J_1 and J_2, and so the group is very definitely selected. Institutions that desire to eliminate individuals of inferior academic abilty would be justified in denying admission to those who are found in the lowest 10-15 per cent according to tests,—unless these individuals can present evidence of previous work of high quality.

The second low group divided at the lower 40 per cent point is clearly indicated. The rise at 43 indicates that the depression at 20-30 is due to the inclusion of slightly under-mediocre individuals in the lower section according to tests. When the percentage point at which they should have been included is reached, the curve rises. This depression and elevation indicates that the tests reverse the position of some poor and some under-mediocre individuals. There can be but few good individuals within the lower 43 per cent according to tests, for if many good individuals were included in the 20-30 depression the curve could not rise again. The fall in the curve after 43 per cent shows the presence of upper mediocres who cross the midline.

The poor group thus covers the range from 15 to 43 per cent of the lower half of the entire group according to tests. These students are the slower members of the class, and their reaction to the subject matter of the curriculum will not be the reaction of the superior individuals. This group is sharply enough defined to justify the separation of these students from the total group according to their performance in tests.

The most surprising fact shown by the charts is that the lower half of the mixed group may be divided *twice*. This fact is also the most definitely indicated, for not only are the J curves practically identical, but the grade curve of Group G_1 which is subject to many irregular influences in the lowest percentiles shows the same tendency.

The point for division in the upper half of the total group is not so clearly marked. If the J_2 curve were typical we could

make the division at any point up to 25 per cent, for the rank-tangential is well over 80. The J_1 curve is similar to the J_2 curve in shape, although it drops sooner to a considerably lower level. J_2 is probably more nearly what may be expected, for the J_2 test series is the more recent. Practically all the improvement in the J_2 series seems to have been made with reference to the good group. The depression in G_1 at 12, as will be shown later, is probably due in part to the use of grades as the criterion of ability.

The upper division may best be made at about the 15 per cent point. The height of the curves at lower 43 indicates that scarcely any of the high grade individuals can be in either of the groups, and therefore they must be in either the good or the mediocre group. The temporary misplacement of a few good students into the mediocre group should not work seriously to their disadvantage.

The lack of form of the upper judgment curves, J_1 and J_2 of chart I, as contrasted with the lower half of the curves suggests the possibility that failures of the selective agencies (entrance requirements, etc.) may have altered the character of the poorer half of the class. College freshmen as a whole are selected from the more capable of the population at large, and hence any irregularities in the selective agency will make themselves felt principally in the lower half of the freshman group. The tendency will be to make the upper section a more uniform sampling, and the elevations in lower half of chart I are quite intelligible when explained on the basis of faulty selection.

There would be little justification for dividing freshmen classes of much larger size than those here considered on the basis of the points which are indicated by the charts on the basis of this evidence. It seems likely on a priori grounds that the points at which division may be made most accurately are functions of the amount of selection which has previously operated upon the freshman group. This selection is necessarily less rigid on a group of 500 than it is on a group of 50. However, the resemblance between the two curves from groups that might be expected to be quite dissimilar, indicates that small freshman groups may be

very much the same as samples of the general population. The similarity of the curves suggests that if the points of division are carefully determined for a single large freshman class of an institution, successive classes may be divided immediately after testing, at the determined points.

6

The value of the series of mental tests in selecting homogeneous groups of students depends ultimately upon the relationship between standing in tests and academic ability; and so to find a high value for the tests, as we have, is to present direct evidence for a close relation between standing in tests and ability. But there is an indirect way of estimating the relation, namely, by observing the relation between standing in tests and the grade criterion which is known to be a faulty index of ability. If it can be shown that changes in the degree of relationship are just what would be expected if the tests actually were measures of academic ability, the evidence for tests as measures of ability is just that much stronger.

Chart II shows values of t for the relation between performances in tests and grades for groups G_1 and J_1 when the group is divided according to grades.

In general, the relation between test standing and judgments of academic ability is seen to be closer than that between test standings and grades. If tests were really a measure of intelligence, a closer relation with judgments would be expected, for on a priori grounds judgments have been shown to be a better criterion of intelligence than grades.

There are two points of general similarity in the contour of the curves of chart II. (1) In the upper half, there is a tendency to rise from a low rank-tangential to a higher one as more and more individuals are included in the upper group. This tendency stops in J_1 at 28, although in G_1 it continues to 42. This accords with the hypothesis that many high grade minds are content with resting slightly above the average grade, and therefore the highest relation with the tests of ability would not be found until these individuals are included in the group which are called good according to grades. (2) In the lower half, the curves

both show a depression in the middle of the range. The recovery in J_1 is marked; in G_1 it is so slight that it is of little significance. The depressions themselves are interestng however, for they may be due to the fact that many individuals who are capable intellectually get low grades because of social or athletic activity. This being true, a fall in the relation between grades and a measure of intelligence would be expected at the points where these individuals are found.

The differences between the two curves of chart II are partially explained by the difference in the attitude of these two groups of students toward their work. It is unfortunate that the increase of pressure that was put on the students cannot be exactly measured; the influences which were used have already been mentioned in Section III.

There are two principal differences in the curves of the two groups. (1) First and most striking is the great height to which the J curve rises in the lowest section. This is because only the students who get the lowest grades stand lowest in the tests. This is not true for the G group. If it is true that all the students in J are putting forth a great deal of work on the subjects of the curriculum, it would be expected that no students would get low grades except those who are incapable intellectually. (2) The depression in the middle half of the lower section is nearer the centre of the total group in the J curve. If one of the effects of the more serious attitude of the J group were to prevent those in J_1 who neglect their work from being content to drop so low as those of G_1, and if the series of mental tests were actually a good method of measuring academic ability, the shift in the depression from a lower point in G_1 to a more central point in J_1 would be expected.

This indirect evidence based upon the discrepancies between the grades of the students and their standing in mental tests although not wholly unambiguous thus tends to confirm the direct evidence given from the close relation between judgments of the academic ability of the students and their standing in mental tests, that performance in mental tests is a general indication of academic ability.

V

To summarize the results of this research:

(1) Methods of estimating academic ability were examined, and the Pearson scale of intelligence was adapted so that it might be used in securing estimations of students' abilities from instructors.

(2) Two series of superior tests judged by their correlations with estimated ability were selected for evaluation with respect to their success in selecting groups of individuals, homogeneous in academic ability, from the total freshman class.

(3) A formula has been deduced to express the extent to which a certain number of individuals at either extreme of a variant Y holds that average rank position in variate X. The coefficient resulting from the use of this formula has been called the rank-tangential coefficient.

(4) The series of tests were evaluated, and it was found that they may be used with considerable accuracy in selecting groups of students of homogeneous academic ability.

(5) The points at which division may best be made were determined for freshman classes of small size, and a method was devised whereby points of division may be found for any series of tests and for groups of any size. Small groups seem best divided at the lower 10-15, the lower 40-43, and the upper 15 per cent points. The two lower points are definitely indicated; the upper point is less clearly shown, although the most recent series of tests gives a very high correlation with estimated academic ability when an upper group including the best 25 per cent according to tests, is separated from the entire class.

(6) Indirect evidence for a close relationship between test standing and academic ability was found from the changing values of the rank-tangential coefficient as the group is divided at different points.

The value of mental tests is not confined to dividing mixed classes into homogeneous groups just because they will do this work successfully. In many institutions it may be impossible or undesirable to conduct separate classes for different grades of ability. In such cases the standings in the tests may be used

simply to tell in which of the groups the student would have been placed in case the divisions had actually been made. With this information the student's advisor may talk more understandingly with the student about his difficulties, and the administrator may regulate the election of courses and the amount of extra-curricular activity of an individual with greater confidence. The more extended usefulness of mental tests in academic work should not be forgotten in the consideration of the work of separation which they do so well.

STANDING IN TESTS. J_1

150	135	120	105	90	75	60	45	30	15	— 0 +	15	30	45	60	75	90	105	120	135	150	
																					460
																					445
										1											430
																					415
										1											400
																		1			385
											1			1			2				370
													2	1							355
					1	1	1			1	1		1				1				340
										4	1	2				1					325
				1		1		2	1	1	3		1								310
						1		1			1		2								295
							1		2												280
								2	1												265
3		1																			250
				1																	235
																					220
			1																		205
																					190
																					175

Estimated Ability

STANDING IN TESTS. J_2

150	135	120	105	90	75	60	45	30	15	— 0 +	15	30	45	00	75	90	105	120	135	150	
																					460
																	1				445
																					430
																					415
															1						400
								1													385
									1			2	1			1	1				370
										1		1									355
									2		2				1						340
				1						1	1	1									325
								2		2	1										310
							1	2	2	2	1										295
	1				1			1													280
			1					1													265
	1																				250
																					235
									1												220
																					205
																					190
																					175

Estimated Ability

APPENDIX I

Description of Tests

Hard Directions. For J_1 the Woodworth Hard Directions[29] form without the final heavy bar was used. Because of ambiguity, the first instruction was cancelled, so that the test began, "Put a comma etc."

For J_2 a form of the test modified by Professor Woodworth was used. The changes were as follows. In line 2 the digits 2, 4, 6, 8, 9 were substituted for the letters F G H I J. In line 9 the space between *sentence:* and *"A horse* was reduced. At the end of line 22 the word *or* was added.

A form which may better be used embodies further changes which seemed necessary. In line 8, the phrase beginning *Put in a number* was changed to *Put the correct number in the next sentence.* It was necessary to introduce the idea that the number to be supplied was the right number, and that it was to be put into the incomplete sentence, not directly after the colon.

The scoring formula used was Index: 3 t + e where t and e are scores in time and errors respectively expressed as deviations from the mean in terms of the standard deviations as the unit deviation. Approximately this formula was given by the regression equation for both groups.

In the directions to the subject, one should emphasize that both speed and accuracy will be measured. It is also advisable to instruct the subject not to stop after the test is begun.

The test is given to subjects individually.

Absurdities. The absurdities test was constructed along the plan outlined by Simpson.[20] The following ten sentences are printed on strips of cardboard and are placed before the subject in a pile. Five of the sentences are logical and five are absurd.

1. Having reached the goal, I looked back, and saw my opponents still running in the distance.

2. The storm which began yesterday morning has continued without intermission for three days.

3. While sharpening his three bladed knife, my cousin cut his second finger.

4. Phyllis was born three years before her younger sister, Ruth.

5. Our office boy may get ahead at last; he was often behind before because his watch was slow, but he has been coming early of late.

6. Preferring a tarnished reputation to the possibility of becoming a corpse for the rest of his life, the young soldier took to flight.

7. Mrs. Smythe has had no children, and I understand the same was true of her mother.

8. Having dressed carefully and elaborately, she descended to the breakfast room, only to find it deserted.

9. The hands of the clock were turned back, so that the time of the sun's rising might seem later.

10. That day we came in sight of several icebergs that had been entirely melted by the warmth of the Gulf Stream.

Sentences 1, 2, 6, 8, 10 are Simpsons. Sentences 3, 5, 9, are adaptations, and sentences 4 and 7 are new. The form of the test here described is that used with J_2 with the following exceptions. In sentence 3, the word *second* has been substituted for *middle,* and sentences 7 and 9 have been interchanged.

In the sentences given to J_1, one was too easy and another was ambiguous. The sentences have been replaced by others, and consequently the means and standard deviations of the two groups are not comparable.

In order to determine whether sentences were ambiguous in their absurdity, the subjects were asked to find their mistakes after the test proper was concluded. In no cases did subjects feel that their error was due to the possibility of more than one interpretation of a sentence.

The score is measured in time and errors. No consistent scoring formula was found for this test, although in both cases time was found to be a better measure of ability than accuracy. For J_1 an index was used, $I = 3\,t + e$, but for J_2 time was used alone since the formula obtained was $I = 23\,t + e$. It is not possible to make a recommendation for a scoring formula at this time,

but there will probably be little lost if the score is measured in time alone.

The instructions to the subject are as follows: On each of these ten cards there is a sentence. Some of these sentences are logical and others of them are absurd. An example of an absurd sentence is the following: I have three brothers, Paul, Henry and myself. (The absurdity is explained if necessary). I want you to tell me which of the sentences are sensible and which are absurd. Don't look for mistakes in spelling and in grammar, but for *logical* absurdities. If a sentence is all right, say *"Right."* If it is absurd, say *"Absurd."* As soon as you have decided about a sentence turn the card over and tell me about the next one.

It has been found that any suggestion concerning speed or a stop watch will produce an undesirable state of excitement in some subjects. If the test is given near the end of a series of speed tests, the subject will not be likely to waste any time during the progress of the test.

The test is given to subjects individually. Since the sentences are not of equal difficulty, the test is hardly adapted for group work.

Sentences Built. Three words, *citizen, horse, decree,* were given to the subjects, with the instructions to construct as many sentences as possible from these. A time limit of five minutes was set. The score in the test was the number of sentences constructed. If the last sentence was only partially complete, it was counted as a complete sentence. It is recommended that the word *petition* be substituted for the word *decree.* Occasionally a subject will not know the meaning of the latter word.

This test was given as a group test.

Opposites A. The words used in this test were

long	dead	east
soft	hot	day
white	asleep	yes
far	lost	wrong
up	wet	empty
smooth	high	top
early	dirty	

In timing the test, the watch is started with the first response. No scoring formula can be recommended. For J_1, index $= 2$ t + e was used; for J_2 time alone was taken.

This test is given to subjects individually. The responses are oral.

Opposite B. This opposites test consists of a series of words of slightly greater difficulty than list A. The procedure is exactly the same. The words:

good	push
north	over
heavy	young
less	city
sharp	wild
sick	rich
big	open
weak	war
come	sell
male	innocent

In view of the recent work of King & Gold[11] on this test, no recommendations are made concerning the content of the lists.

The score in this test was time alone.

Analogies A and B. This is the mixed relations test described by Woodworth.[28] Series A is the card beginning *good: bad:: long* —Series B; *eye: see:: ear*—.

The score for these tests was time of the interval between the first and last responses.

This test is given to subjects individually.

8. Alphabet. The alphabet or alphabet sorting test was used by Burt.[3] The materials required are two complete alphabets of letters, each letter printed on a single card. The size of the card is one square inch. The letters may be readily obtained in a game called *Anagrams* published by Parker Brothers, Salem, Mass. The letters are then numbered from 1 to 52 on the backs of the cards, and the same order is always used. This order is:

Z P Q K T C M Z T L H R E O B L
W I K P Y B V A D N E J U R C G S V
F U X J S A W Y X G N I D Q F H M.

The letters are covered by a cardboard screen, and the subject is seated before it. The letters are not exposed until the test begins.

Each subject is given the following instructions orally: Under this screen there are two complete alphabets of letters—that is 52 letters—all mixed up. The two alphabets are just alike. Now I want you to pick up letter A, and put it here; (experimenter indicates position) then take letter B and put it beside letter A; then letter C; and so on until you have one complete alphabet. You must be sure to pick up letters in alphabetical order; you mustn't pick up letter H until letter G is placed. I want you to do this as fast as you can, and I think you can do it faster if you arrange the alphabet in two rows;—go from A to M in the first row, and from N to Z in the second,—so that you won't have to reach so far. Use both hands if you wish.

The screen is removed, and as soon as the experimenter sees that it is no longer obstructing the vision of the subject, the stopwatch is started. The watch can be carried in the right hand while the screen is being removed, and thus the time of starting may be secured with great accuracy. The watch is stopped as soon as the letter Z is placed in position.

If subjects ask questions before the experiment begins, answers are given. Errors occur rarely, and no penalty is given on this account. Subjects must be held strictly to picking up the letters in alphabetical order.

9, 10, 11, 12. Logical Memory Tests. The passages used in these tests are given in *The Scientific Study of the College Student*. For J_1, the procedure followed, and the scoring system adopted, were exactly those described by Dr. Kitson. For J_2 the tests were given as group tests, but the scoring was changed so that in the Auditory Memory test the credit given for the main sections was 5, 10, 10, 25 units respectively, with partial credit for partial reproduction. In the Visual Memory Test, the credit for the main sections was 5, 15, 15, 15. Each idea reproduced was given 1-3 unit; and an individual's total score was found by adding the units credited for ideas to the units credited for the principal sections of the passages.

Neither these passages nor this scoring formula will probably be found in the final statement of the logical memory tests. The indications are that credit should be given for the actual number of words written, instead of for the main passages. The evidence is that each word written should receive half the credit that is given for an idea correctly reproduced. However, any scoring formula that could be recommended would be practically an arbitrary one. The absence of an ultimate scoring formula should not cause the abandoning of the tests, for even with the tests in this crude form good results were secured.

For J_2, the length of time between the first and the deferred reproduction was changed to three weeks. The results were very unsatisfactory. If the deferred logical memory tests are used, the period between tests should not be greater than two weeks.

APPENDIX II

Pearson[15] has deduced a formula which offers a method of expressing the relation between a measured variable, and a second variable which is divided at some point into two classes. The meaning of the formula may be better understood from the diagram. It is assumed that the second variable would show a Gaussian distribution if measurements were made, and that the regression of X on Y is linear. When the means of the two variables are taken as o,

$y = b_1 x$, where b_1 is the regression-coefficient of X on Y. That is, $\frac{y}{\sigma_y} = r \frac{x}{\sigma_x}$.

The problem is to evaluate x and y in the diagram. x is

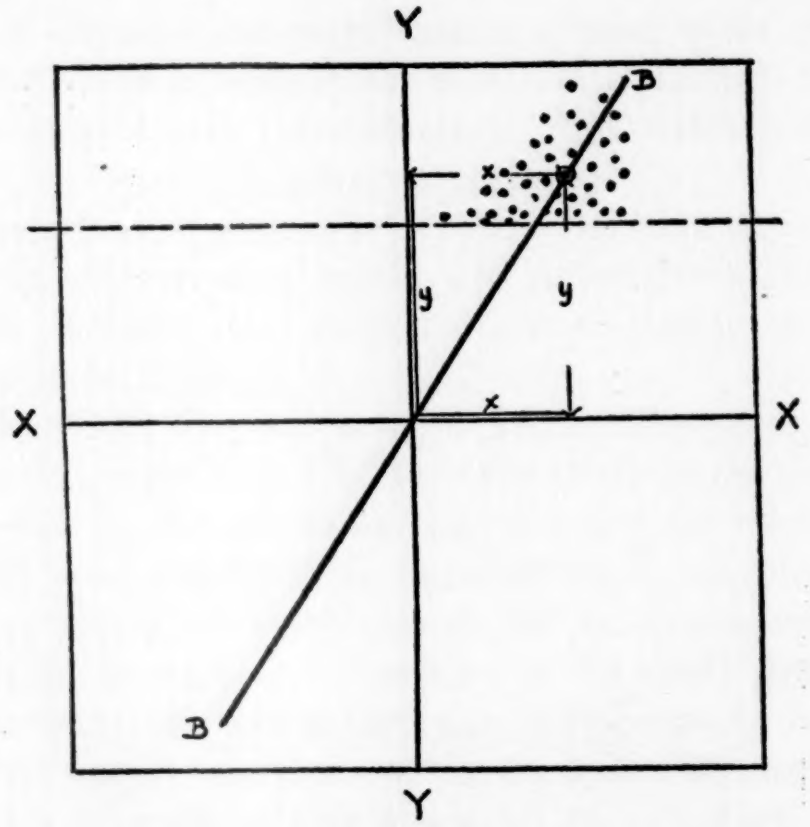

Diagram I.

easily found; it is the mean value of the measured variable which is found in one of the classes of the alternative variable. Y is

the distance from the centroid of the class of Y in which the average of the measured variable was taken, to the mean of Y. The assumption of the Gaussian frequency for the alternative variable makes possible the computation of this distance from a knowledge of the proportion of the alternative variable which is included in the class in question.

$$\frac{y}{\sigma_y} \text{ thus equals } \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2}}{\frac{1}{\sqrt{2\pi}} \int_{\frac{y}{\sigma}}^{\infty} e^{-\frac{1}{2}y^2} dy}$$

The steps in the computation of r by this formula are given elsewhere.[18]

If both variables are given in actual measurements, the limitation of Gaussian distribution in the alternative variate is removed, since for the value of $\frac{y}{\sigma_y}$ we may use the observed value for the distribution of the alternative variate with which we are dealing. However, if the distribution is closely Gaussian, the original formula may be used with considerable saving of time.

The second assumption made by Pearson is that the regression of X on Y is linear. Then "if a volume of the frequency be cut off from the frequency surface by a vertical plane at a given value of the variate Y, the vertical through the centroid of this volume cuts the regression line." If r is to have the same value for any point at which the frequency surface may be cut by the vertical plane, the vertical through the centroid of this volume must cut the regression line for any of these points of division. But this latter condition can be met only when the mean value of X associated with every Y array however small also lies on the regression line. Such strict linearity is rarely found, even though the correlation table as a whole may seem to indicate clearly a linear regression. Consequently, a certain fluctuation in the values given by the formula for different points of division of the Y variate is to be expected. *The psychological interest lies in determining whether the fluctuations are the same for differ-*

ent samplings of the same population, and in discovering what causes may lie behind them.

The actual relation between the two variables, X and Y, does not change no matter where we may divide the frequency surface, and therefore it seems confusing to call coefficients obtained by this method *coefficients of correlation.* It is conceivable that a regression might test linear according to the Blakeman[2] criterion, and yet hardly a single coefficient computed by this method be equal to the product-moment coefficient. For this reason it is suggested that the coefficient be called the tangential coefficient, and that it be designated by the letter T. The name tangential is selected since the coefficient is the slope of the line connecting the vertical through the centroid of the solid cut from the frequency surface, with the axis along which the Y variate is represented (when the variates are measured in terms of their standard deviations). T is equal to r when the vertical through the centroid of the solid cut off cuts the straight line which best fits the means of the X's associated with Y arrays. It is equivalent to r in meaning since it gives the mean value of X deviations which are associated with the mean of a Y array.

T is a good measure of relationship between the two variables, since high values of T indicate that the average of X variations associated with a Y array is near the mean of the Y array, and when the regression is roughly linear this can happen only through a close relation between the variables. But when the regression is markedly skewed, T may exceed + or —1 for very small selected classes, and thus fail to give any idea of relationship.

It is also apparent that T is not exactly suited to our conditions, since it depends for its value upon a ratio of *measurements* in *both* variables. Now clearly in the variable given by alternative categories, we are not interested in measurements. We assign a certain percentage of the total group to one class, and the rest of the group to the other. Thus we need only know *which* individuals are included in each of the alternative categories *regardless* of what their relative positions may be. However, if we wish to change the size of the selected group, it is necessary

to know the *order of merit* of the individuals in the variable according to which selection is to be made; for in the first division we may take the superior 10, in the second division the superior 12 etc. *Consequently, our purposes are fulfilled if we but know the rank position of each individual in the variable given by alternative categories. A completely satisfactory coefficient will take this fact into account.*

In the case of the continuous variable, the measurements themselves might be used; yet, for the sake of uniformity with the alternative variable, and in order to attain the greatest simplicity in computation, it seems desirable to use rank position in the continuous variable as well.

The Rank-Tangential Coefficient

To overcome the difficulties of the tangential coefficient and to provide an index of relationship more in harmony with the concrete situation, a formula has been deduced to express the extent to which a certain number of individuals at either extreme of the Y variate holds that average rank position in the X variate. The relation between the variables is thus measured in terms of likeness in the ranks of the individuals instead of similarity in actual measurement.

A coefficent parallel to the tangential coefficient would be given by expressing the ratio of the mean value of X deviations in rank associated with a Y array, to the mean rank deviation of that Y array. See diagram. Since the number of cases in X equals the number in Y, and since measurement is in terms of rank, the standard deviations of the variables are equal, and may be neglected.

Let the number of individuals upon which the measurement of the relation between the two variables is based be N; and let the variable Y be divided so that M individuals at one extreme of the distribution are separated from the remaining N—M.

The mean of each variable is $\frac{N+1}{2}$; and the mean of the Y array is $\frac{M+1}{2}$. The line y in the diagram is therefore equal to

$\frac{N+1}{2} - \frac{M+1}{2}$; that is $\frac{N-M}{2}$. *This is the Y deviation.*

The rth X deviation is equal to $\frac{N+1}{2} - X_r$. Therefore the mean of the X deviations found in the Y array of M cases must be $\frac{N+1}{2} - \frac{\Sigma(R_x)}{M}$. *This is the mean rank deviation for the X's in the Y array.*

The required ratio is

$$\frac{\frac{N+1}{2} - \frac{\Sigma(R_x)}{M}}{\frac{N-M}{2}} = \frac{M(N+1) - 2\Sigma(R_x)}{M(N-M)} = t$$

where $\Sigma(R_x)$ is the sum of the ranks in variable X of the M best or worst individuals according to variable Y.

For the sake of simplicity of treatment, the individual who makes the best performance in a test is given rank 1, regardless of whether the numerical value of a superior performance is relatively high or low.

The sign of the denominator depends upon the particular extreme, the good or the poor, of the distribution which is under consideration, and not upon the numerical values of N and M. For example, if the superior individuals of the Y variate are taken as M, y is a positive deviation, and the denominator is positive. If the inferior individuals of Y are considered, y is a negative deviation, and the denominator is negative.

It is suggested that t be called the rank-tangential coefficient. t proves to be a good measure of relationship. Its meaning is definite and readily understood; it varies between $+1$ and -1; its value is zero if the two variables are independent; if $M = N$, t is indeterminate. Finally, the coefficient is easily computed.

The following illustrations show how t is computed.

I. Let $N = 10$; $M = 4$. Divide according to the Y variate, selecting the four superior cases. Suppose the ranks of these

four cases in X are 1, 3, 4, and 6. Substitute in formula N = 10; M = 4; S(x) = 14

$$\frac{4(10+1)-2:14}{4(10-4)} = +.66$$

The denominator is positive since M includes the superior cases of Y.

II. Let N = 20; M = 12. Let M include the inferior cases, and let their ranks be 6, 7, 8, 9, 11, 12, 13, 14, 16, 17, 19, and 20. S(x) = 152.

$$\frac{12(20+1)-2:152}{12(20-12)} = \frac{-54}{-96} = +.542$$

The denominator is negative since M includes the 12 inferior cases of Y.

The rank-tangential coefficient is not a method of approximating the product-moment coefficient; it is a method of evaluating a relation of another sort. To the writer's knowledge the only method which has been used previously in evaluating such relationships is that of stating the percentage of identical cases, a method which is inadequate because of its ambiguity. It is only because the rank-tangential coefficient does give an index of this type of relationship that the writer feels justified in introducing it into a literature already crowded with coefficients of one sort or another.

The rank-tangential coefficient has properties similar to the product-moment coefficient of correlation:

(1) The rank-tangential coefficient gives the mean of X deviations associated with the mean of Y arrays. In this special case, however, there are only two arrays.

(2) The rank-tangential coefficient varies from +1 for a complete relation between the two variables, through 0 for independence to —1 for complete dissimilarity.

The writer is not able to state the probable error of the rank-tangential coefficient. Judging from the probable error of the bi-serial correlation coefficient (which is equivalent to the tangential coefficient in the case of a normal correlation surface) the probable error of the rank-tangential coefficient will be from

20 to 200 per cent larger than the probable error of the product-moment coefficient. The probable error of the bi-serial correlation coefficient increases very rapidly as one group decreases so that it includes less than 10 per cent of the total frequency.[31]

REFERENCES

[1] Bell, J. C. Mental Tests and College Freshmen. J. Educational Psychology. 1916. VII, p. 381.

[2] Blakeman, J. On Tests for Linearity of Regression. Biometrika. 1905. IV, p. 332.

[3] Burt, C. Experimental Tests of General Intelligence. British Journal of Psychology. 1909. III, p. 94.

[4] Carey, N. Factors in the Mental Processes of School Children. British J. of Psychology. 1916. VIII, p. 170.

[5] Cattell, J. M. American Men of Science.

[6] Dressler, F. B. Psychology of Touch. American Journal of Psychology. 1895. VI, p. 343.

[7] Galton, F. Hereditary Genius.

[8] Gauss, C. F. Methods des Moindres Carres. Trans. par J. Bertrand. 1855.

[9] Gilbert. Mental and Physical Development of School Children. Yale Studies. 1894. II, p. 40.

[10] Hollingworth, H. L. Vocational Psychology. 1916.

[11] King, I. and Gold, H. A Tentative Standardization of Certain Opposites Tests. J. Educational Psychology. 1916. VII, p. 459.

[12] Kirkpatrick, E. A. Individual Tests of School Children. Psychological Review. 1900. VII, p. 274.

[13] Kitson, H. D. Scientific Study of the College Student. Psychological Monographs. XXIII, No. 1. 1917. Whole No. 98.

[14] Merriman, M. Theory of Least Squares. 1894.

[15] Pearson, K. Relationship of Intelligence to Size and Shape of Head. Biometrika. 1906. V, p. 105.

[16] Pearson, K. A New Method of Determining Correlation. Biometrika. 1909. VII, p. 96.

[17] Rowland, E. and Lowden, G. Psychological Tests at Reed College. J. Experimental Psychology. 1916. I, p. 211.

[18] Ruml, B. Measurement of the Efficiency of Mental Tests. Psychological Review. 1916. XXIII, p. 501.

[19] Ruml, B. On the Computation of the Standard Deviation. Psychological Bulletin. 1916. XIII, p. 444.

[20] Simpson, B. R. Correlation of Abilities. 1912.

[21] Spearman, C. General Intelligence. American J. of Psychology. 1904. XV, p. 201.

[22] Stern, W. Differentielle Psychologie. 1911.

[23] Thorndike, E. L. Mental and Social Measurements. 1913.

[24] Thorndike, E. L. Combining Incomplete Judgments of Relative Position. J. Phil. Psych. and Scientific Method. 1916. XIII, p. 197.

[25] Waite, H. Estimations of the General Intelligence of School Children. Biometrika. 1911. VIII, p. 79.

[26] Webb, E. Character and Intelligence. Brit. J. of Psych. Monograph Supplements. 1915. I, No. 3.

[27] Wissler, C. Correlation of Mental and Physical Tests. Psychological Monographs I. 1901. No. 3.

[28] Woodworth, R. S. Combining the Results of Several Tests. Psychological Review. 1912. XIX, p. 97.

[29] Woodworth, R. S. and Wells, F. L. Association Tests. Psychological Monographs. 1911. XIII, No. 5.

[30] Yule, G. U. Introduction to the Theory of Statistics.

[31] Soper, H. E. Probable Error of the Bi-serial Correlation Coefficient. Biometrika. 1915. X, p. 384.